Linear Regression with One Variable

전 재 욱

Embedded System 연구실
성균관대학교

Outline

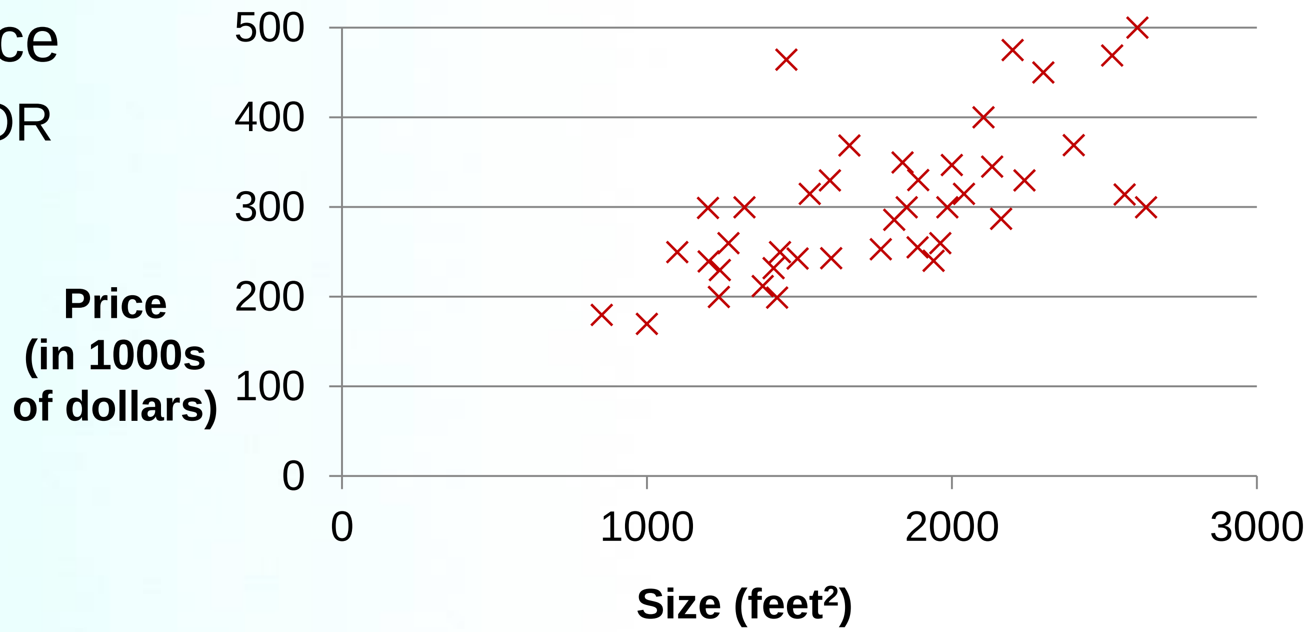
- Model representation
- Cost function
  - Cost function
  - Hypothesis of one parameters
  - Hypothesis of two parameters
- Gradient descent
- Gradient descent for linear regression

Outline

- Model representation
- Cost function
  - Cost function
  - Hypothesis of one parameters
  - Hypothesis of two parameters
- Gradient descent
- Gradient descent for linear regression

Model Representation

- Housing Price
  - Portland, OR



- Supervised Learning
  - "Right answers" are given for every examples in the data

- Regression
  - Predict continuous real-valued output (price)

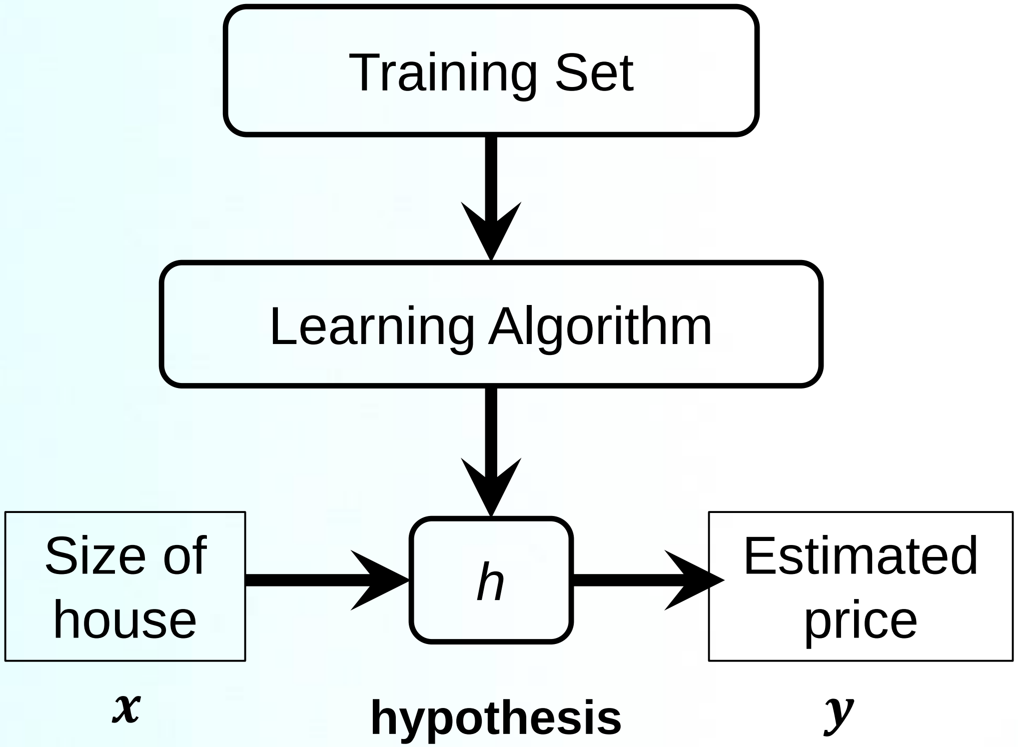
Model Representation

- Training set of housing prices

Size in feet^2 (x)	Size in 坪	Price ($) in 1000's (y)
2104	59.12	460
1416	39.79	232
1534	43.11	315
852	23.94	178
…		…

- Notation
  - m: Number of training examples
  - x's: "input" variable / features, $x^{(1)} = 2104$, $x^{(2)} = 1416$
  - y's: "output" variable / "target" variable , $y^{(1)} = 460$
  - (x, y): one training example, $\left(x^{(i)},\ y^{(i)}\right)$: i-th training example

Model Representation



Model Representation

- How do we represent *h* ?

$$h_\theta(x) = \theta_0 + \theta_1 x$$

  - Linear regression with one variable
    - Univariate linear regression

Outline

- Model representation
- Cost function
  - Cost function
  - Hypothesis of one parameters
  - Hypothesis of two parameters
- Gradient descent
- Gradient descent for linear regression

Model Representation

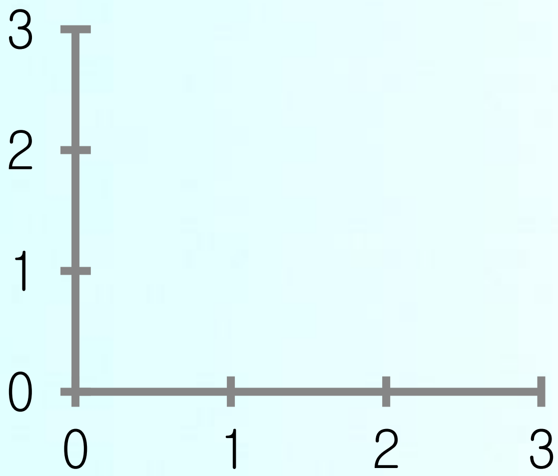
- Training set of housing prices

Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

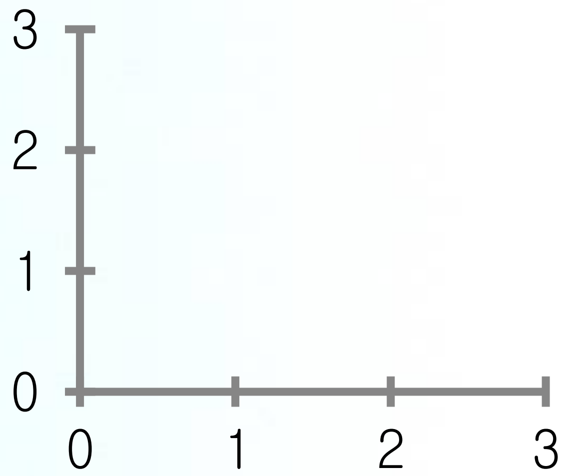
- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$
  - $\theta_i' s$: parameters
    - How to choose $\theta_i' s$?

Model Representation

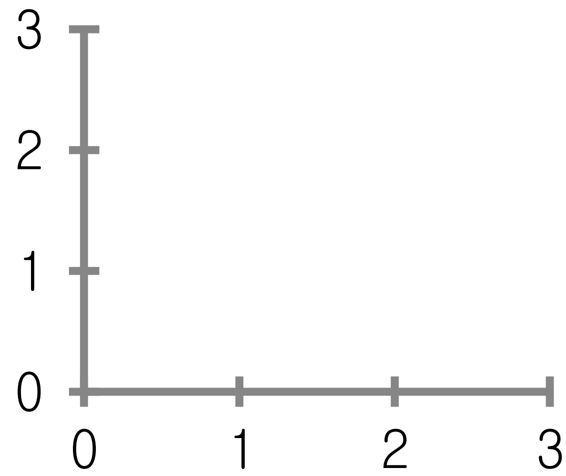
- $h_\theta(x) = \theta_0 + \theta_1 x$



$\theta_0 = 1.5$
$\theta_1 = 0$



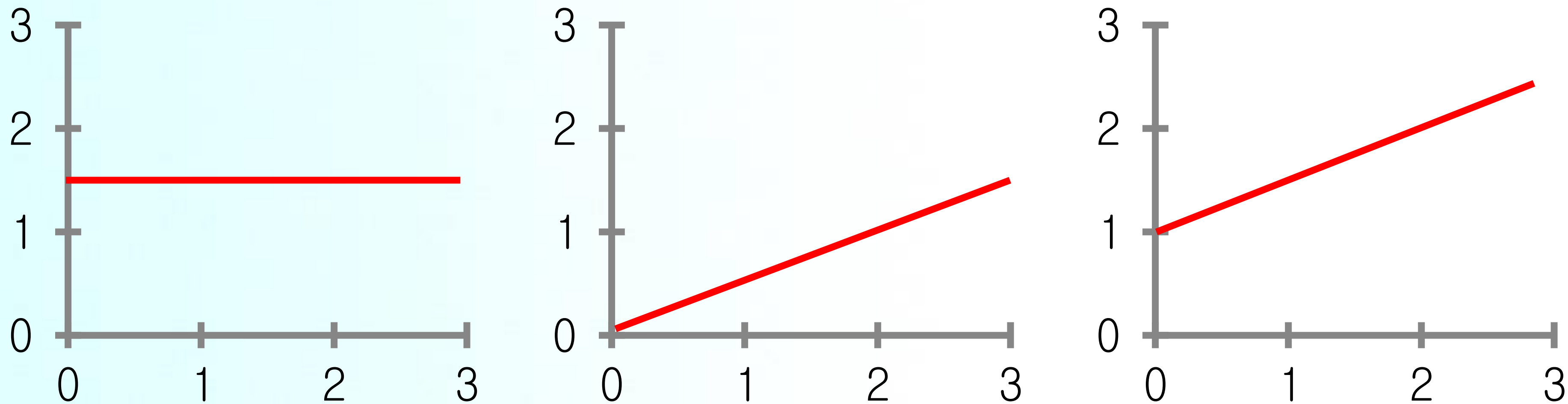
$\theta_0 = 0$
$\theta_1 = 0.5$



$\theta_0 = 1$
$\theta_1 = 0.5$

Model Representation

- $h_\theta(x) = \theta_0 + \theta_1 x$

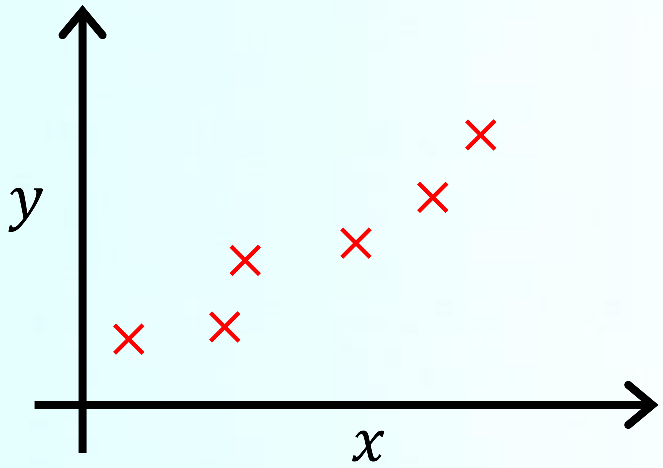


$$\theta_0 = 1.5 \quad \theta_1 = 0$$

$$\theta_0 = 0 \quad \theta_1 = 0.5$$

$$\theta_0 = 1 \quad \theta_1 = 0.5$$

Cost Function



- Idea
  - Choose $\theta_0,\ \theta_1$ so that
    $h_\theta(x)$ is close to y for our training examples (x, y)

$$\min_{\theta_0,\theta_1} \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

- Cost function: $J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$
  - Goal: $\min_{\theta_0,\theta_1} J(\theta_0, \theta_1)$

Outline

- Model representation
- Cost function
  - Cost function
  - Hypothesis of one parameters
  - Hypothesis of two parameters
- Gradient descent
- Gradient descent for linear regression

Simplified Hypothesis

- Hypothesis
  - $h_\theta(x) = \theta_0 + \theta_1 x$
- Parameters
  - θ_0, θ_1
- Cost function

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

- Goal
  - $\min_{\theta_0,\theta_1} J(\theta_0, \theta_1)$

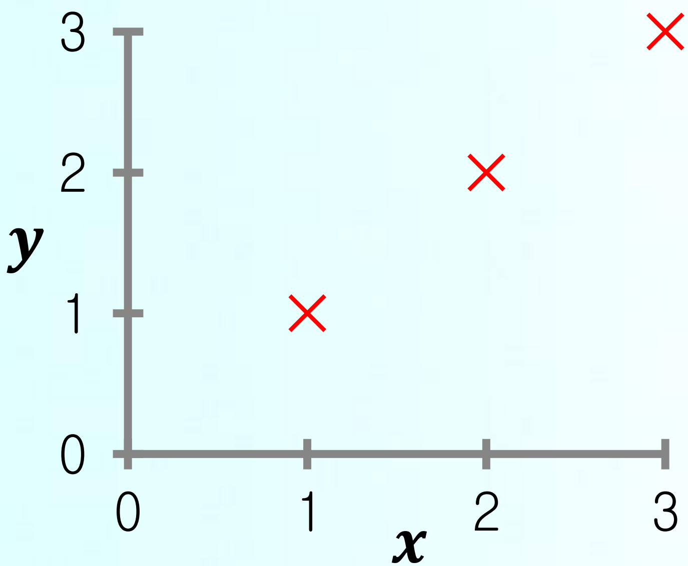
- Simplified Hypothesis
  - $h_\theta(x) = \theta_1 x$
- Parameters
  - θ_1
- Cost function

$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

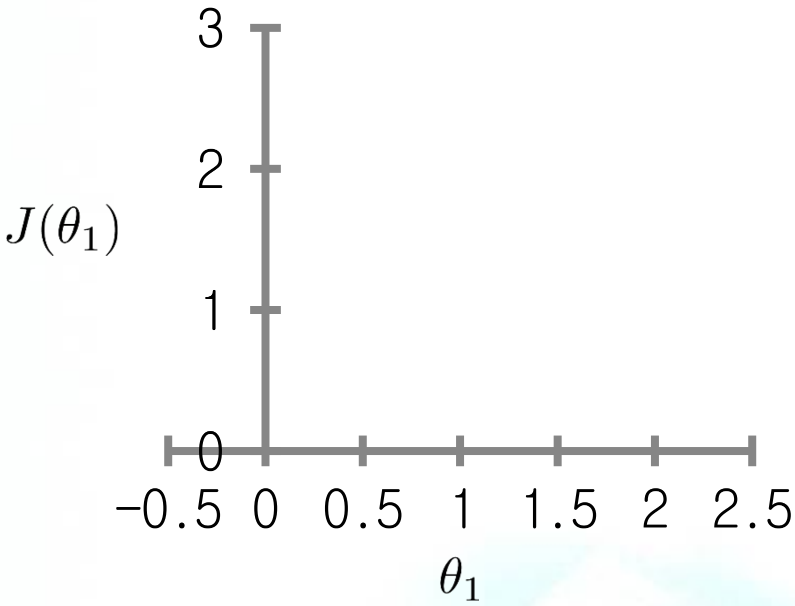
- Goal
  - $\min_{\theta_1} J(\theta_1)$

Simplified Hypothesis

- $h_\theta(x)$
  - a fct of x for fixed θ_1

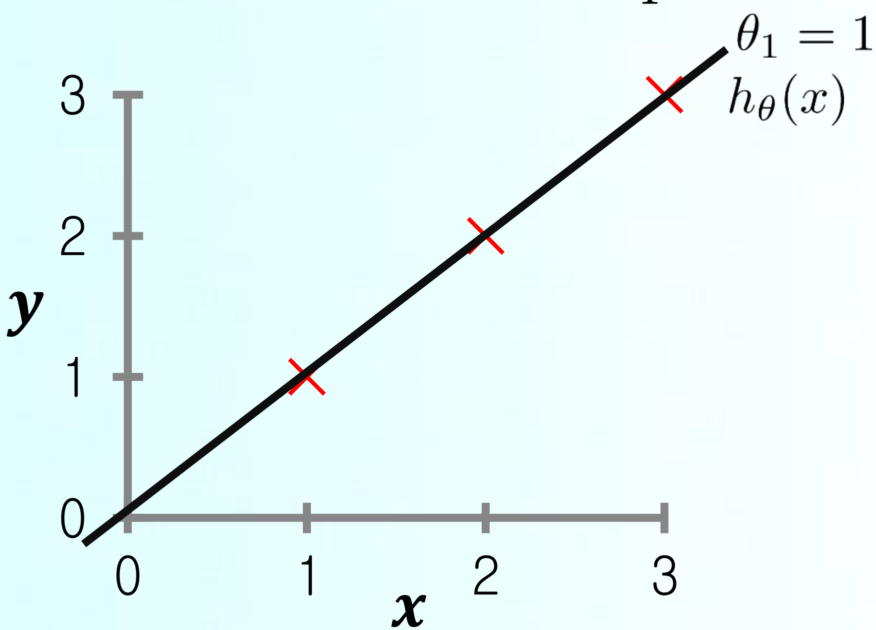


- $J(\theta_1)$
  - fct of the parameter θ_1

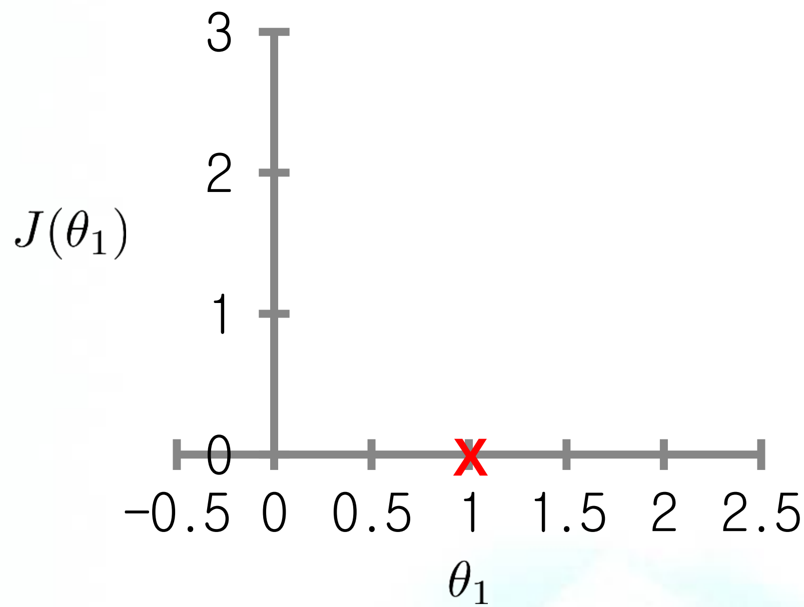


Simplified Hypothesis

- $h_\theta(x)$
  - a fct of x for fixed θ_1



- $J(\theta_1)$
  - fct of the parameter θ_1

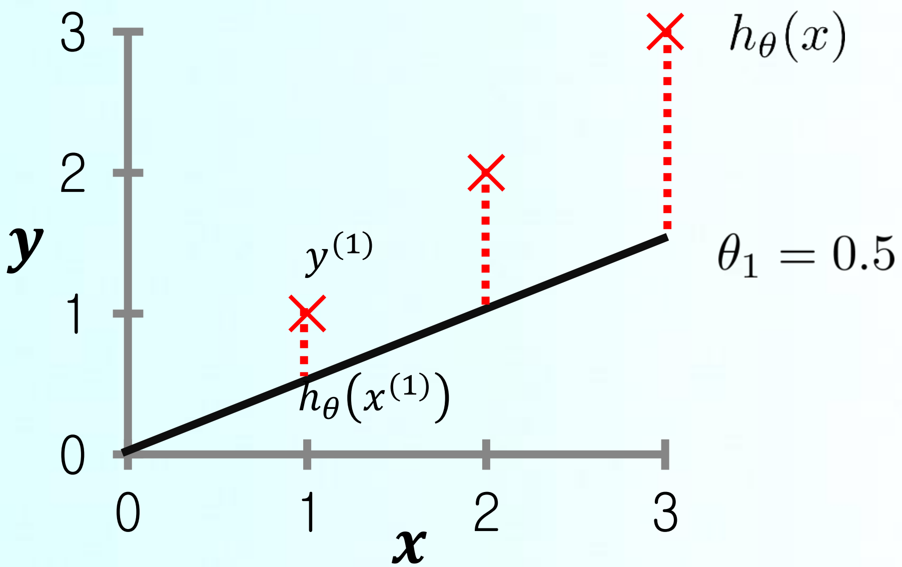


$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$
$$= \frac{1}{2m}\sum_{i=1}^{m}\left(\theta_1 x^{(i)} - y^{(i)}\right)^2$$
$$J(1) = \frac{1}{2m}(0^2 + 0^2 + 0^2) = 0$$

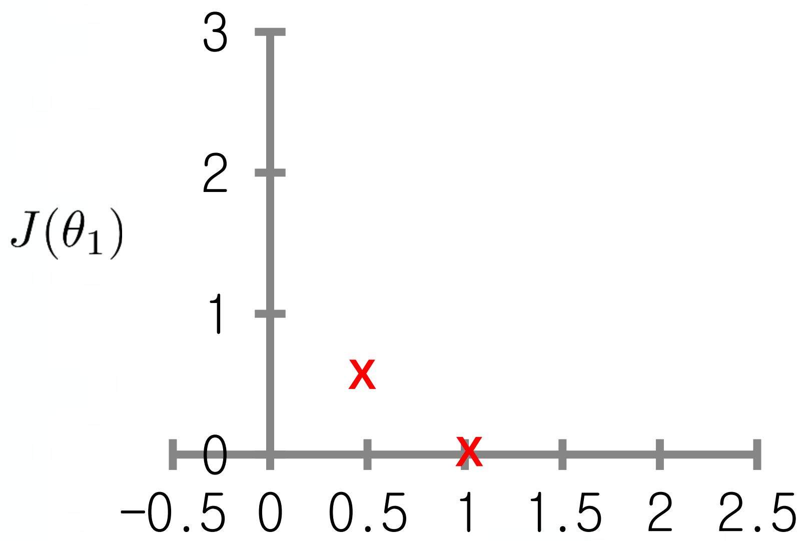
- $J(1) = 0$

Simplified Hypothesis

- $h_\theta(x)$
  - a fct of x for fixed θ_1



- $J(\theta_1)$
  - fct of the parameter θ_1



$J(0.5) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$

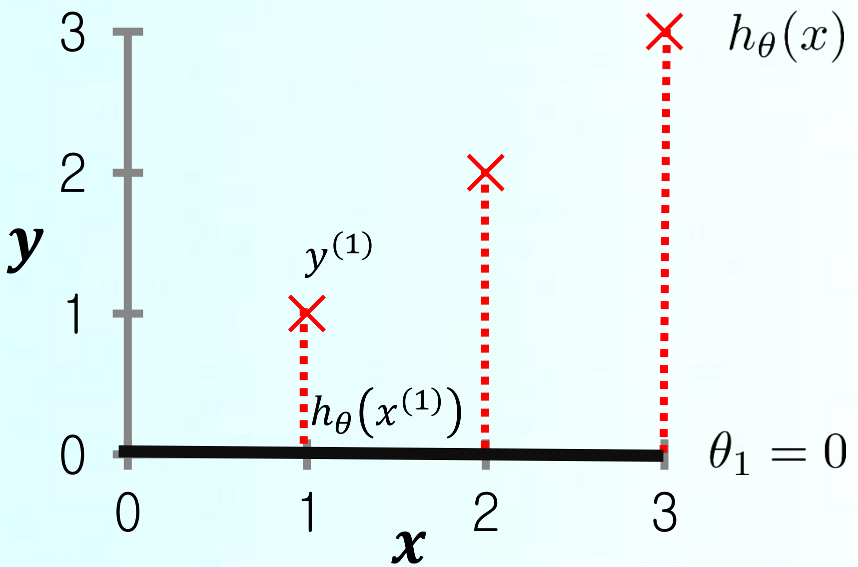
$= \frac{1}{2m}((0.5-1)^2 + (1-2)^2 + (1.5-3)^2)$

$= \frac{1}{2*3}(3.5) \approx 0.58$

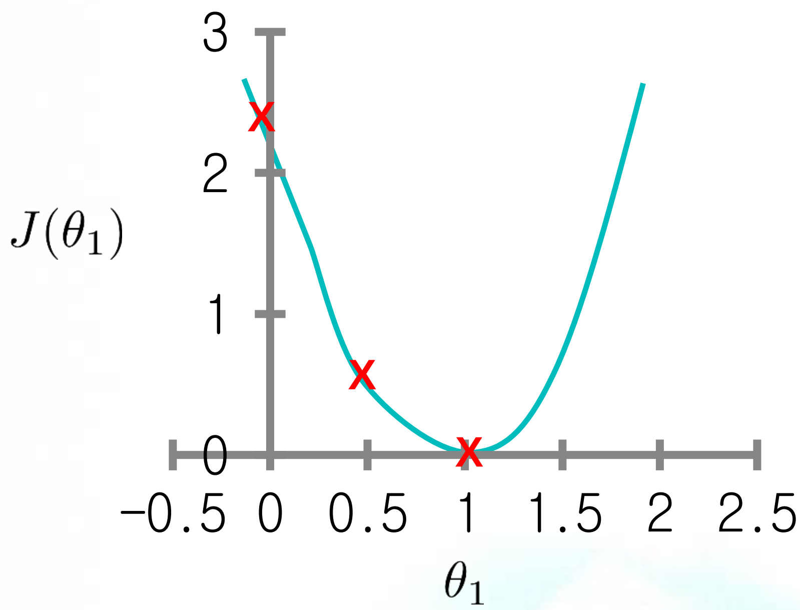
- $J(1) = 0$, $J(0.5) = 0.58$

Simplified Hypothesis

- $h_\theta(x)$
  - a fct of x for fixed θ_1



- $J(\theta_1)$
  - fct of the parameter θ_1



$$J(0) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$
$$= \frac{1}{2m}((0-1)^2 + (0-2)^2 + (0-3)^2)$$
$$= \frac{1}{2*3}(14) \approx 2.3$$

Outline

- Model representation
- Cost function
  - Cost function
  - Hypothesis of one parameters
  - Hypothesis of two parameters
- Gradient descent
- Gradient descent for linear regression

Hypothesis of Two Parameters

- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

- Parameters: θ_0, θ_1

- Cost function

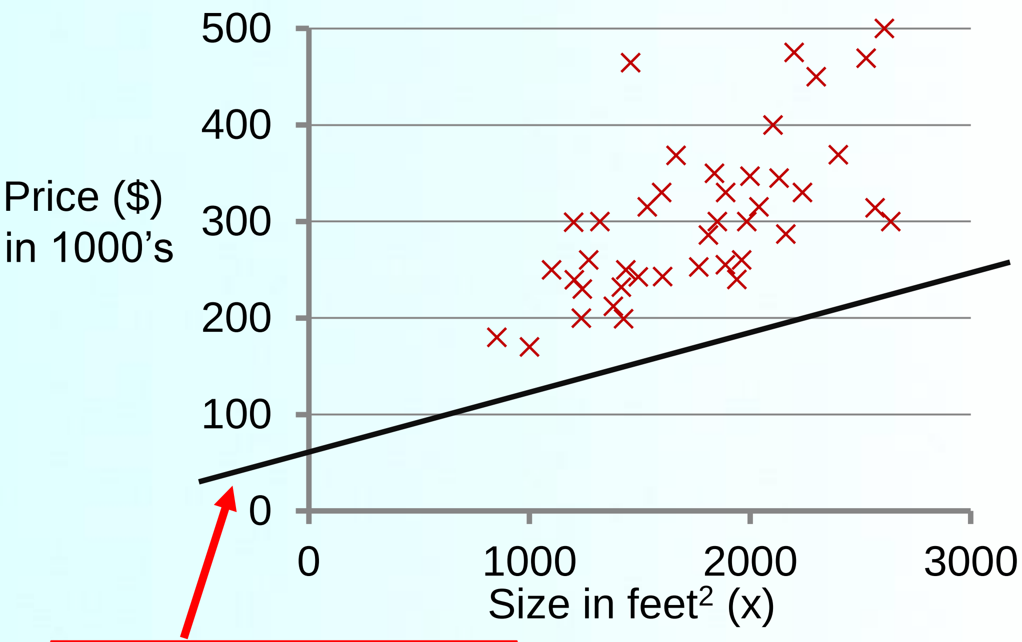
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

- Goal

$$\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$$

Hypothesis of Two Parameters

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1

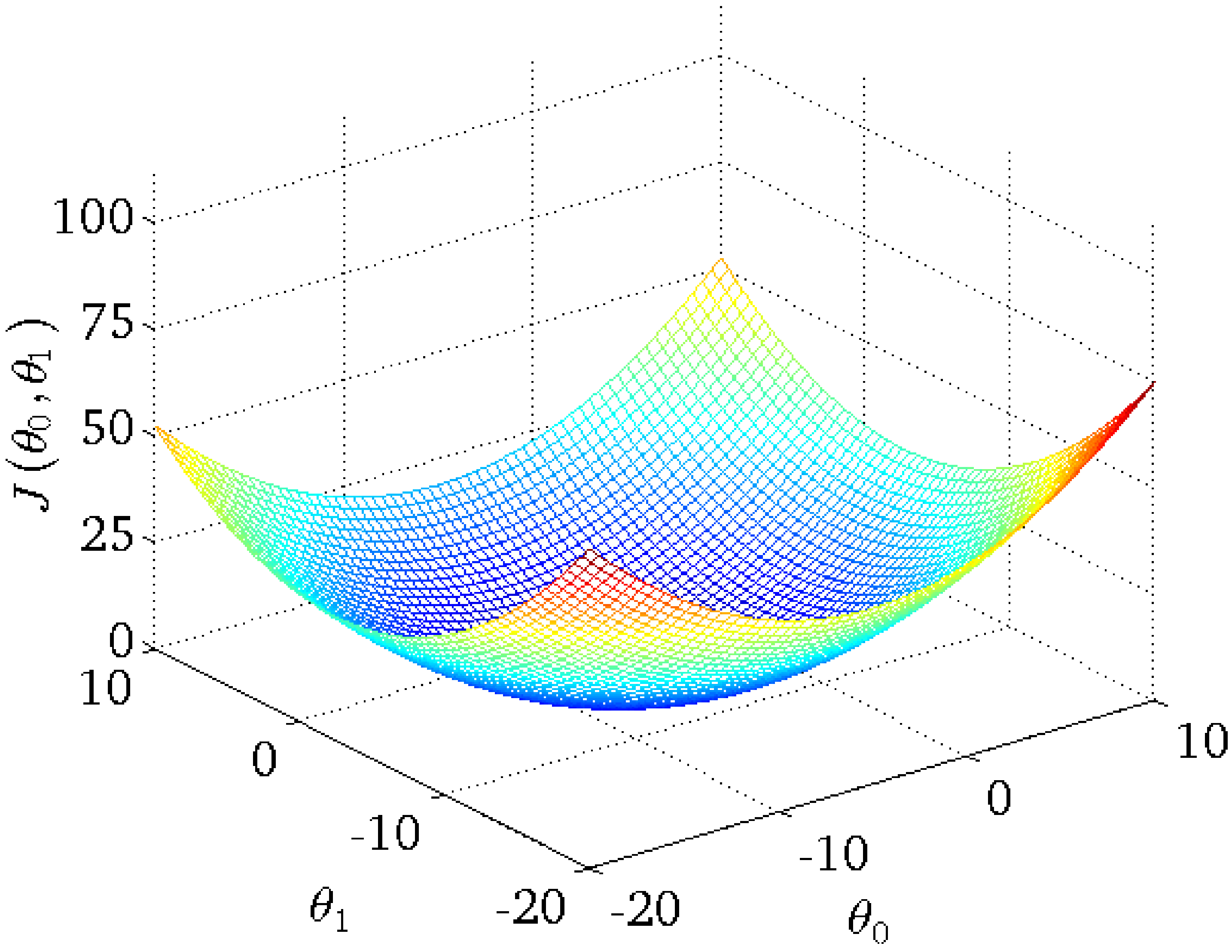


$h_\theta(x) = 50 + 0.06x$

- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1

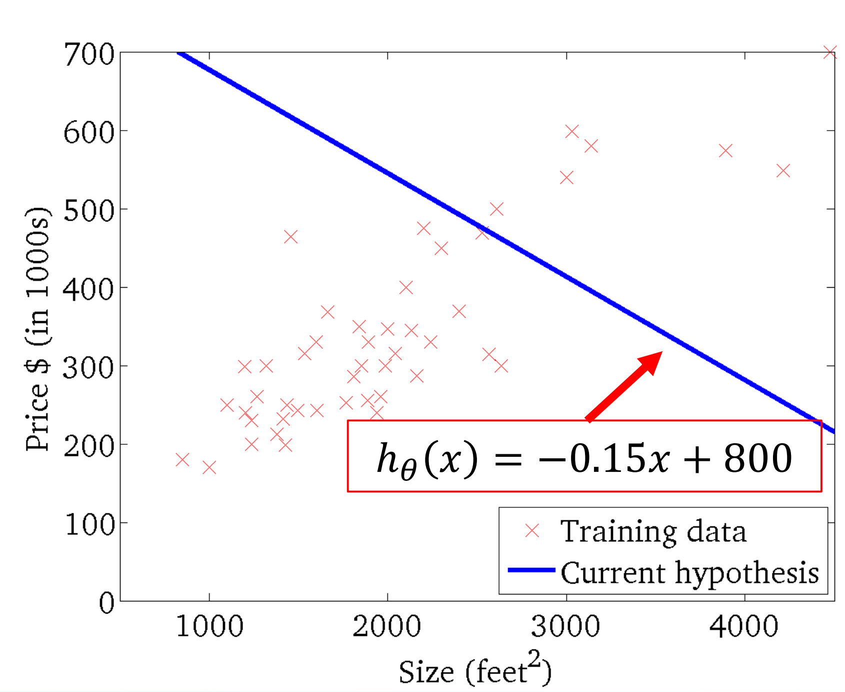
$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Contour Plot

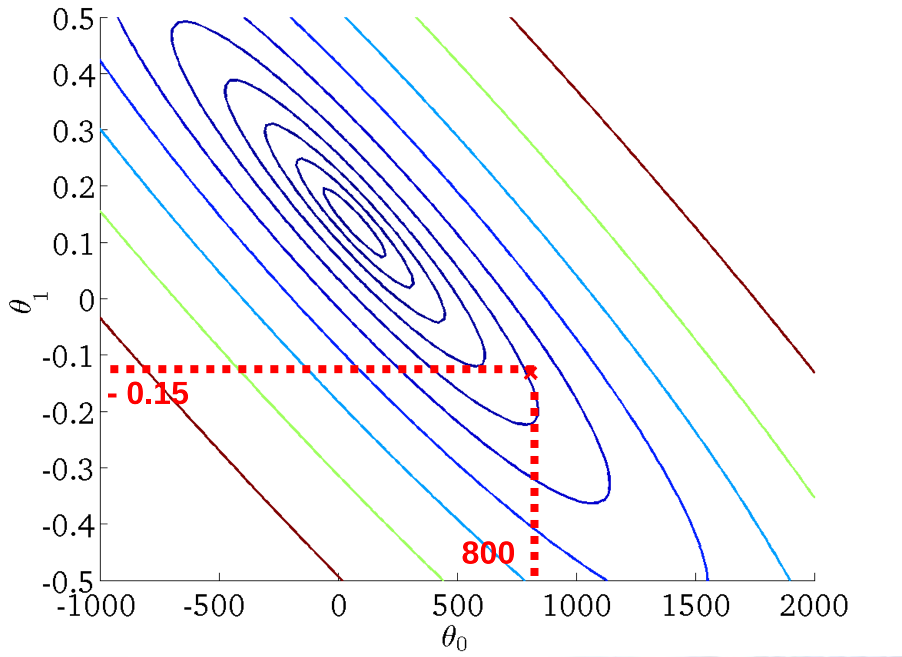


Hypothesis and Its Cost Function

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1

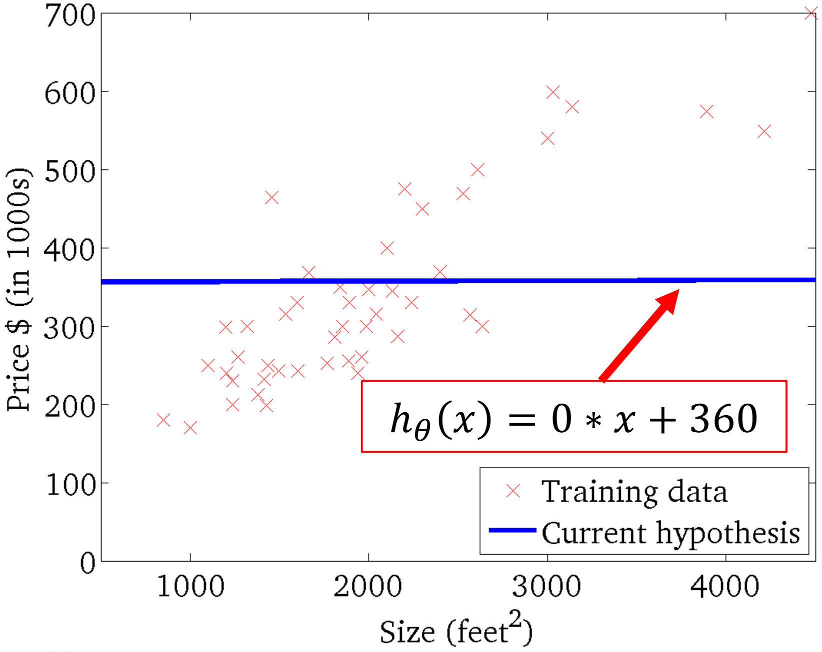


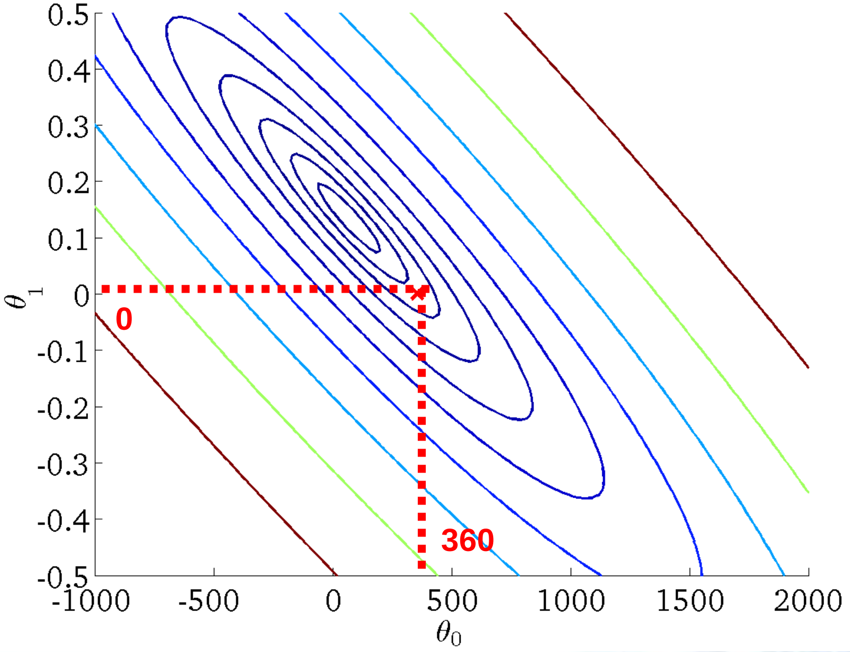
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1



Hypothesis and Its Cost Function

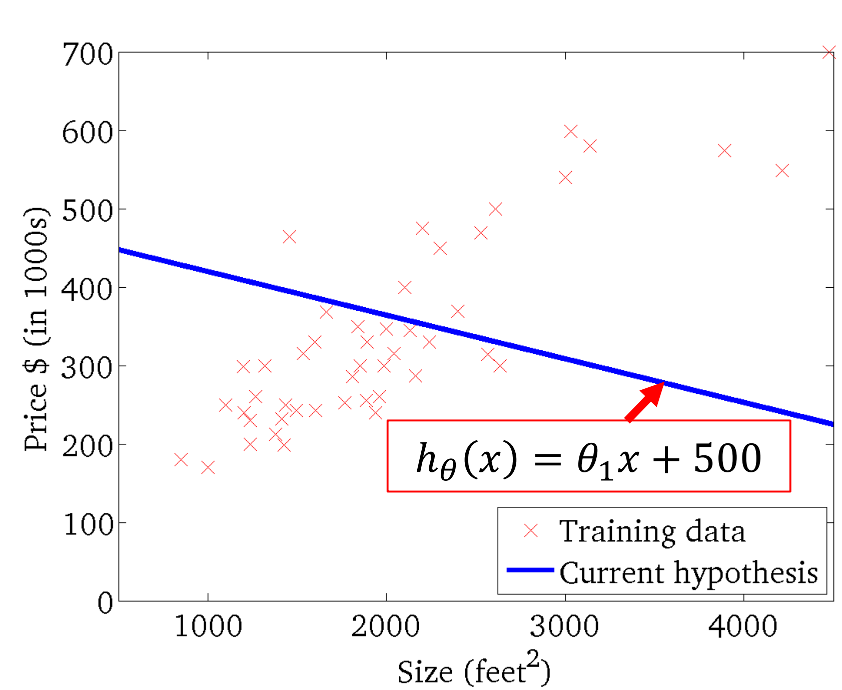
- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1



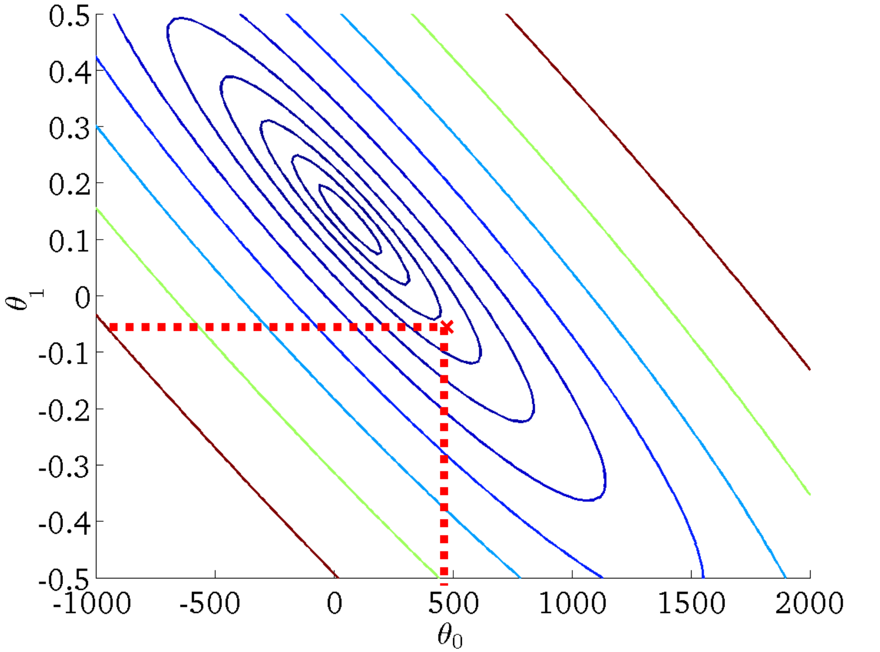


Hypothesis and Its Cost Function

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1
  - $\theta_1 < 0$



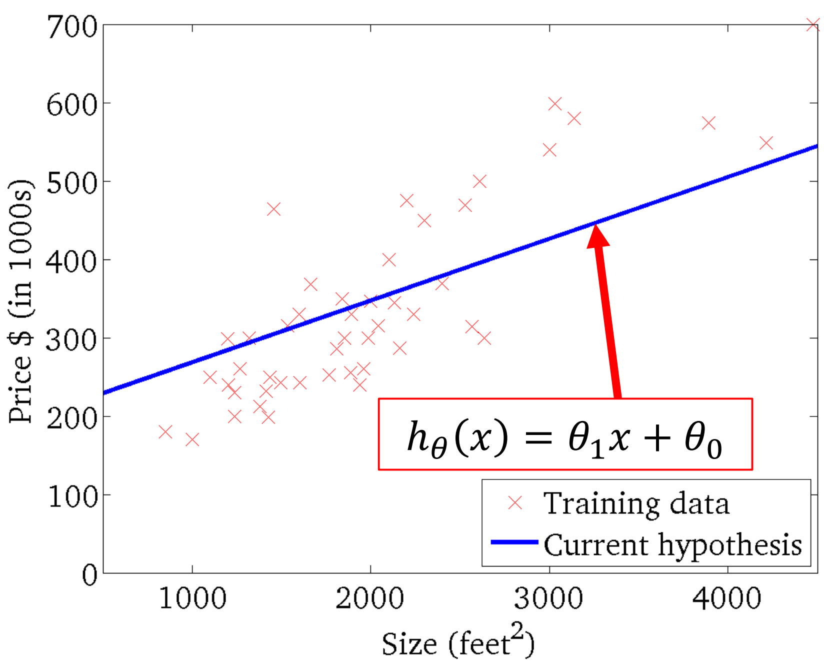
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1

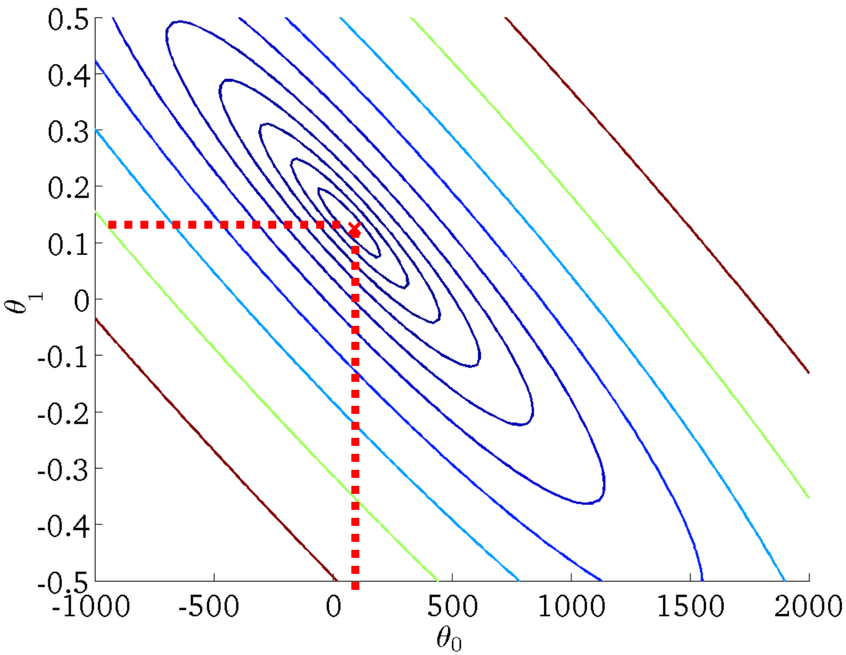


Hypothesis and Its Cost Function

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1
  - $\theta_1 > 0$

- $J(\theta_0, \theta_1)$
  - fct of the parameter $\theta_0,\ \theta_1$





Outline

- Model representation
- Cost function
  - Cost function
  - Hypothesis of one parameters
  - Hypothesis of two parameters
- Gradient descent
- Gradient descent for linear regression

Gradient Descent

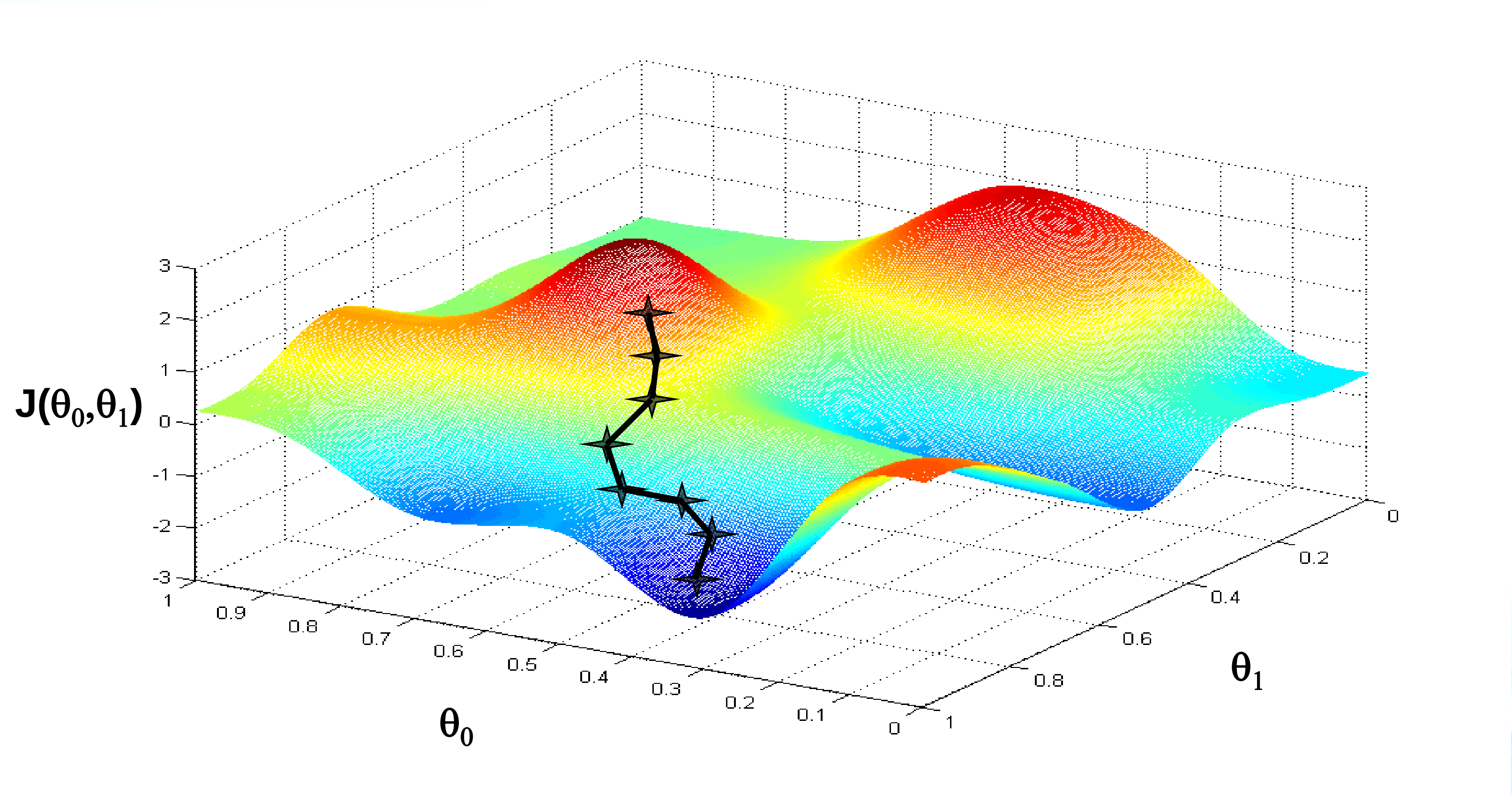
- Given $J(\theta_0, \theta_1)$,

  try to find $min_{\theta_0,\theta_1} J(\theta_0, \theta_1)$

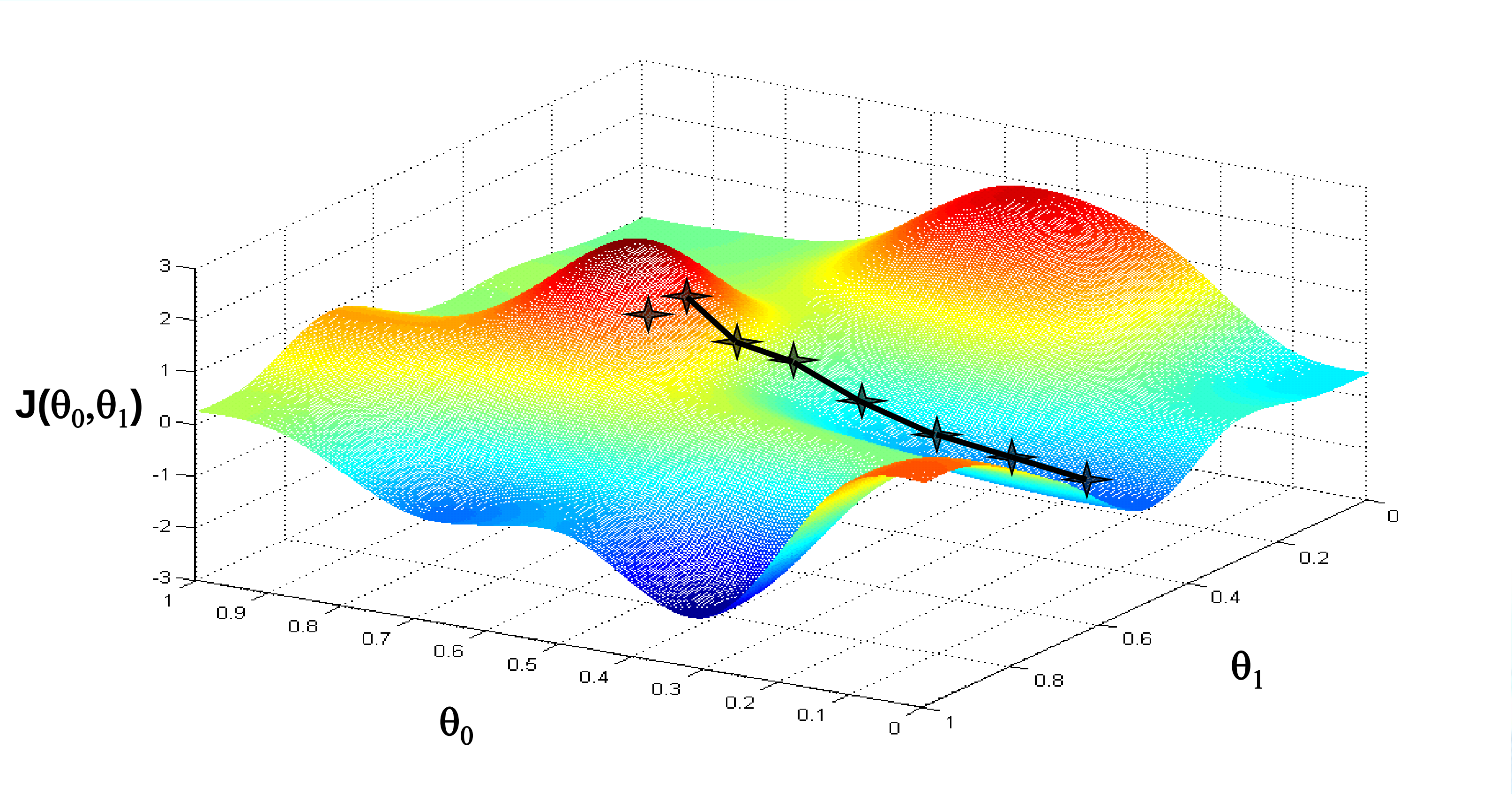
- Outline
  - Start with some θ_0, θ_1
  - Keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$

    until we hopefully end up at a minimum

Gradient Descent



Gradient Descent



Gradient Descent Algorithm

- Repeat until convergence {

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad \text{(for } j = 0 \text{ and } j = 1\text{)}$$

}

- Correct: Simultaneous update

$$temp0 \leftarrow \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$temp1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$

$$\theta_0 \leftarrow temp0$$
$$\theta_1 \leftarrow temp1$$

- Incorrect

$$temp0 \leftarrow \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\theta_0 \leftarrow temp0$$

$$temp1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_1 \leftarrow temp1$$

Linear regression with one variable

- **Gradient descent intuition**

Gradient Descent Algorithm

- Repeat until convergence {

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$ (simultaneously update for $j = 0$ and $j = 1$)

}

Gradient Descent Algorithm

- Repeat until convergence {

$$\theta_j \leftarrow \theta_j - \alpha \boxed{\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)}$$

(simultaneously update for $j = 0$ and $j = 1$)

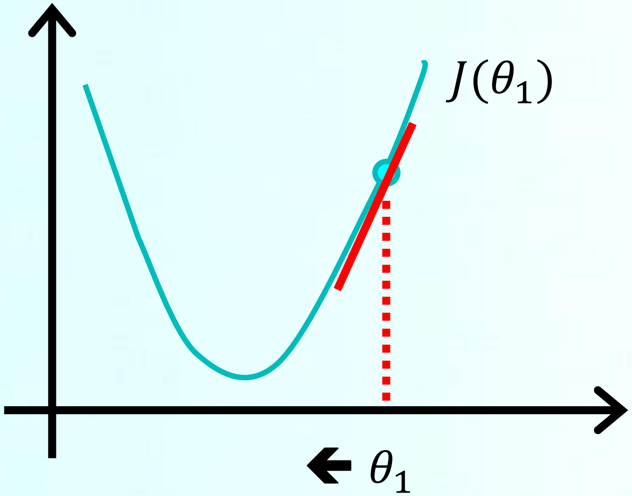
}

Learning rate

derivative

Gradient Descent Algorithm

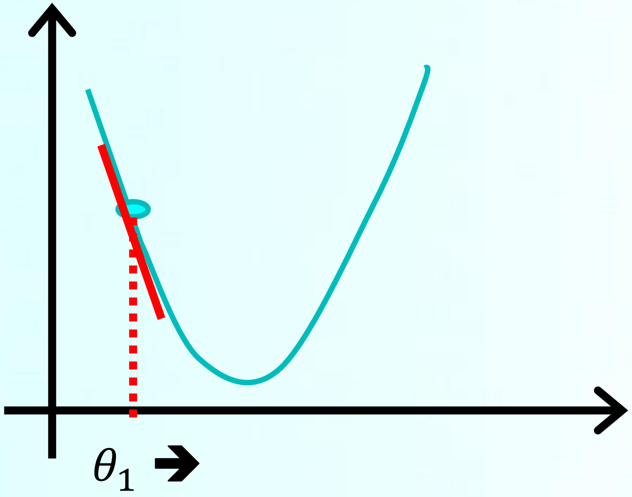
- $J(\theta_1) \quad (\theta_1 \in R)$



$$\theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

$$\frac{\partial}{\partial \theta_1} J(\theta_1) > 0$$

$$\theta_1 \leftarrow \theta_1 - \alpha * (positive\ number)$$



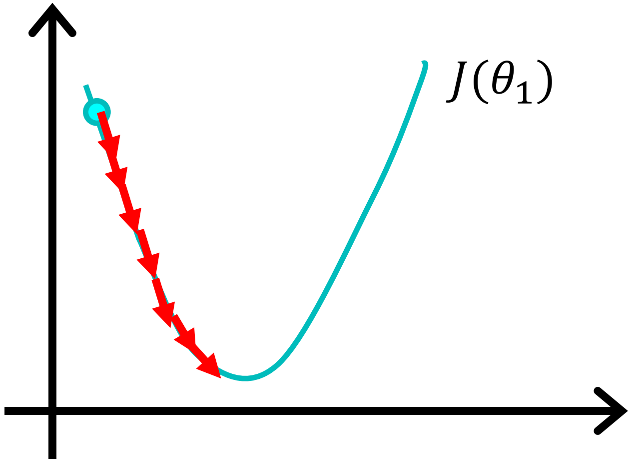
$$\theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

$$\frac{\partial}{\partial \theta_1} J(\theta_1) \leq 0$$

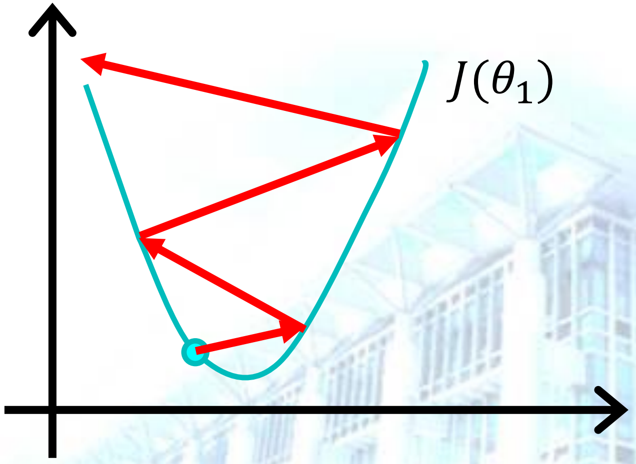
$$\theta_1 \leftarrow \theta_1 - \alpha * (negative\ number)$$

Gradient Descent Algorithm

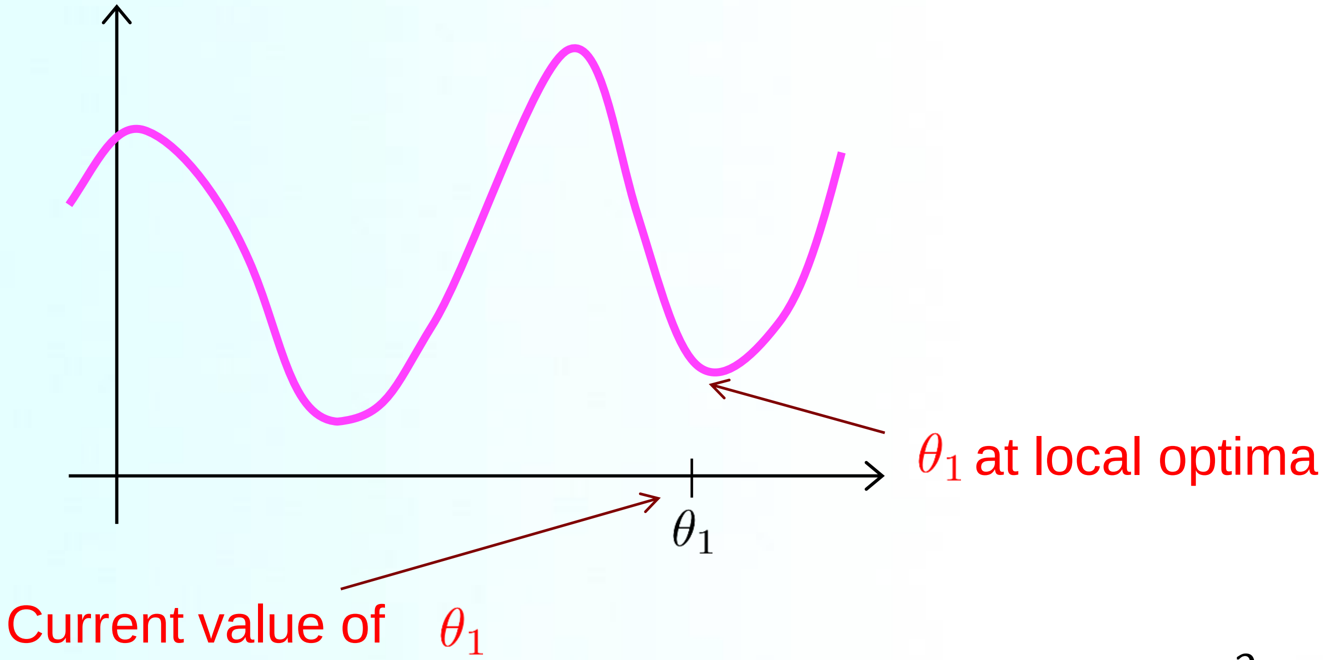
- $\theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$



- If α is too small,
  - gradient descent can be slow.
- If α is too large,
  - gradient descent can overshoot the minimum.
  - It may fail to converge, or even diverge.



Gradient Descent Algorithm



$$\theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

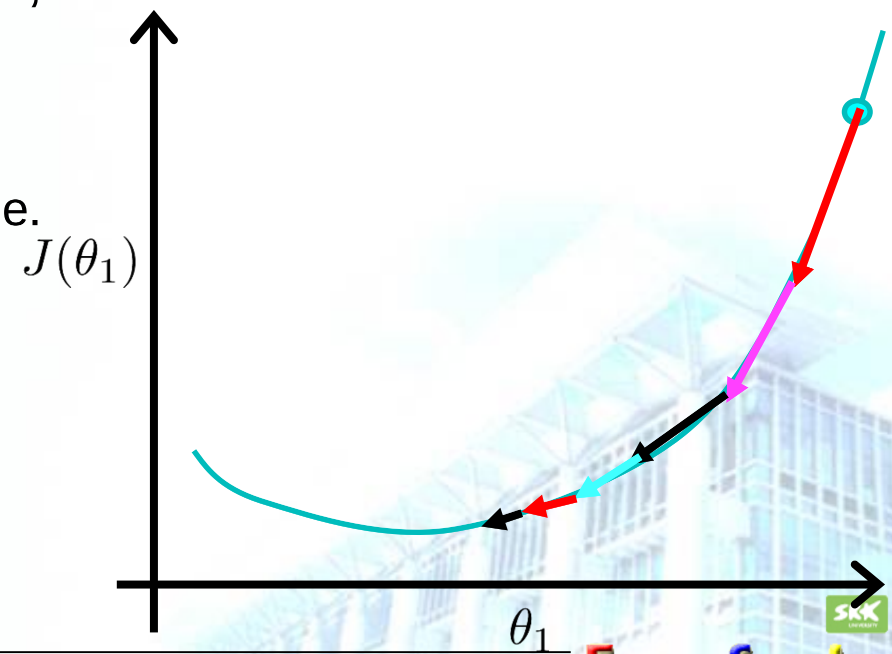
$\theta_1 \leftarrow \theta_1 - \alpha \times 0$ at local min

Gradient Descent Algorithm

- Gradient descent can converge to a local minimum,
  - even with the learning rate α fixed.

$$\theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

- As we approach a local minimum,
  - gradient descent will automatically take smaller steps.
    - So, no need to decrease α over time.



Outline

- Model representation
- Cost function
  - Cost function
  - Hypothesis of one parameters
  - Hypothesis of two parameters
- Gradient descent
- Gradient descent for linear regression

Gradient Descent for Linear Regression

- Gradient descent algorithm

Repeat until convergence {

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

(for $j = 0$ and $j = 1$)

}

- Linear regression model

$$h_\theta(x) = \theta_0 + \theta_1 x$$

- Cost function

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Gradient Descent for Linear Regression

- $\frac{\partial}{\partial\theta_j}J(\theta_0,\theta_1) = \frac{\partial}{\partial\theta_j}\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$

$$= \frac{\partial}{\partial\theta_j}\frac{1}{2m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$$

- $j = 0$

$$\frac{\partial}{\partial\theta_0}J(\theta_0,\theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)$$

- $j = 1$

$$\frac{\partial}{\partial\theta_1}J(\theta_0,\theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x^{(i)} = \frac{1}{m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)x^{(i)}$$

Gradient Descent for Linear Regression

- Repeat until convergence {

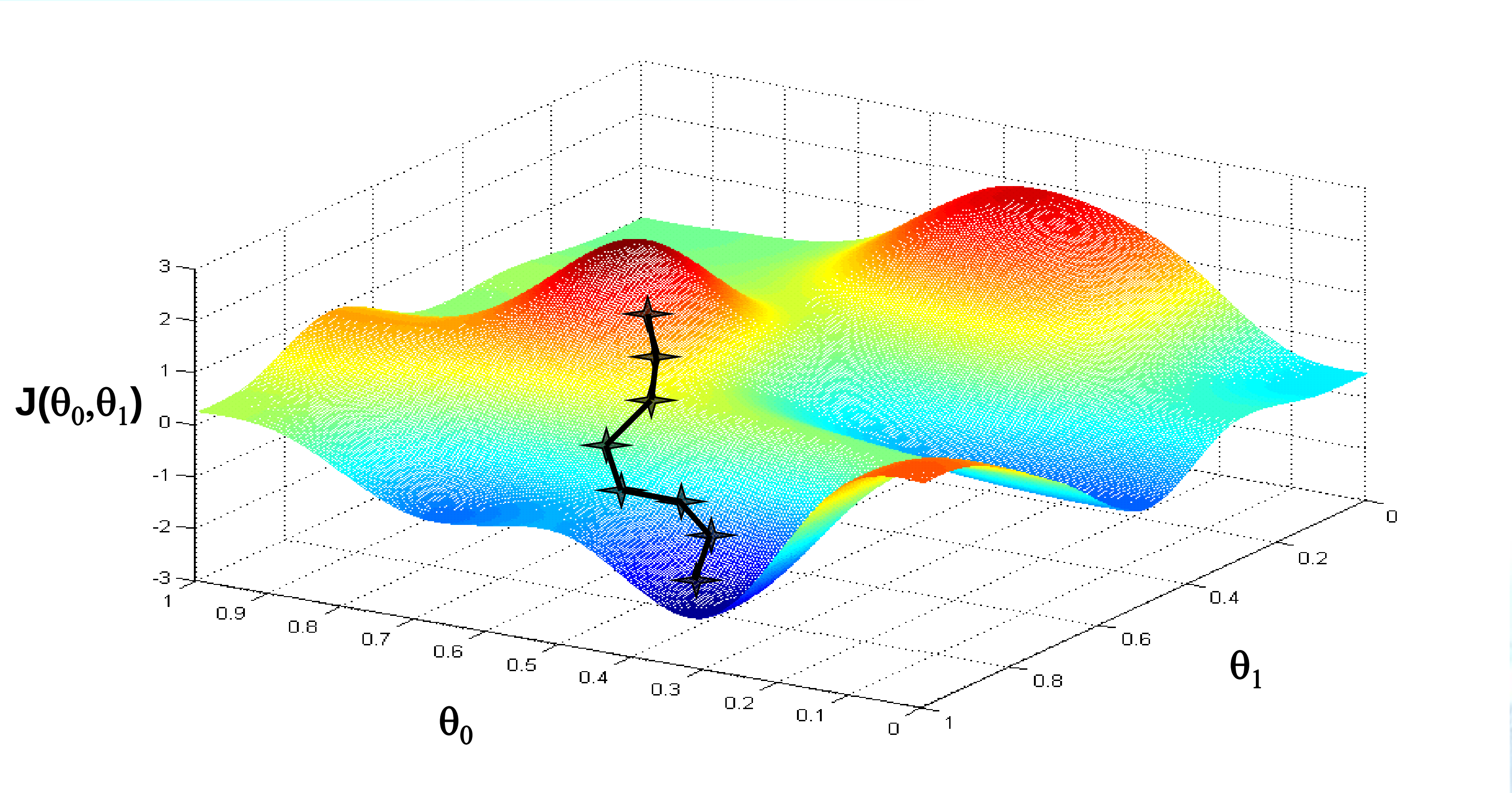
$$\theta_0 \leftarrow \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)$$

$$\theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x^{(i)}$$

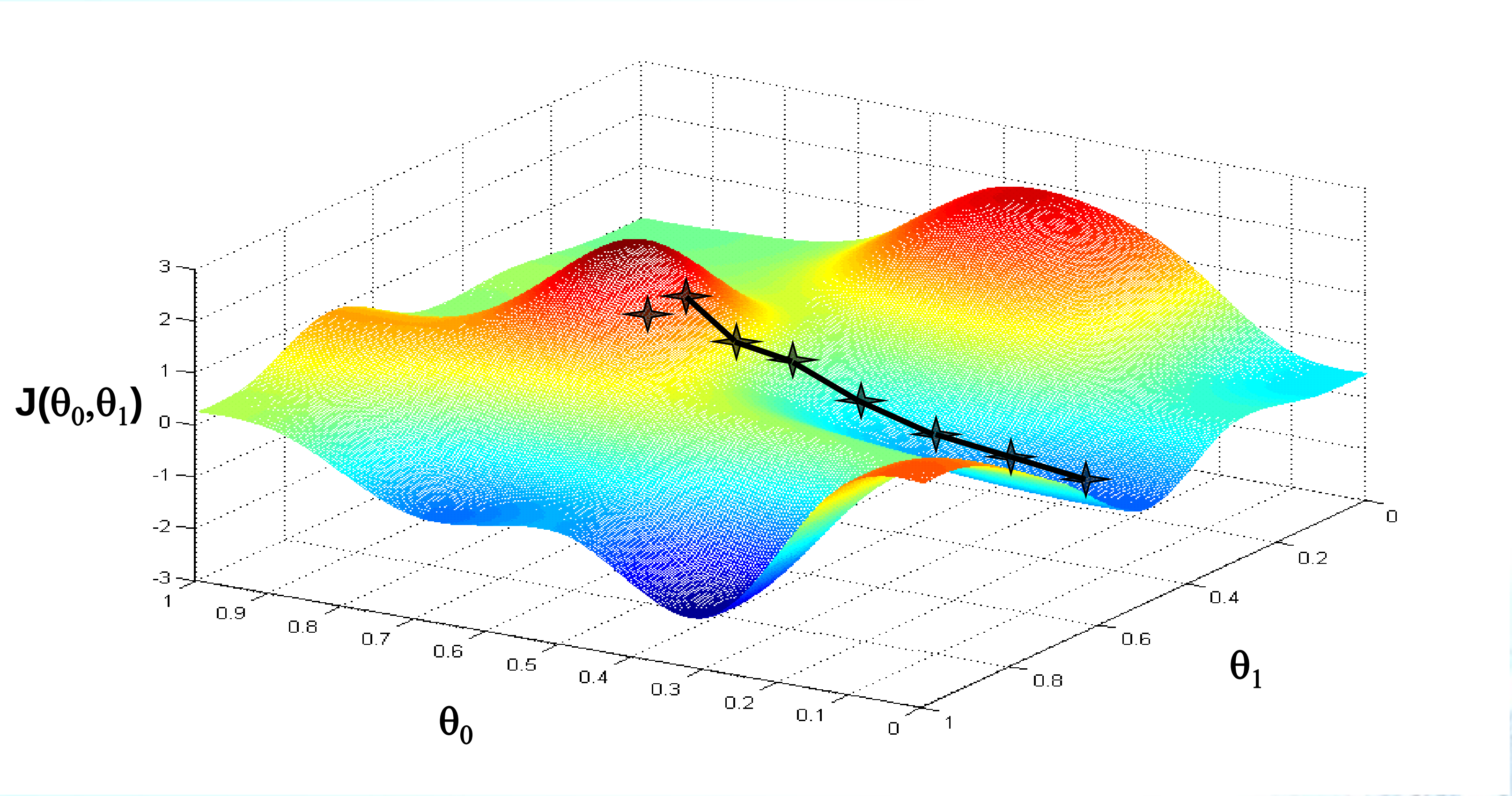
}

(Update for θ_0 and θ_1 simultaneously)

Gradient Descent

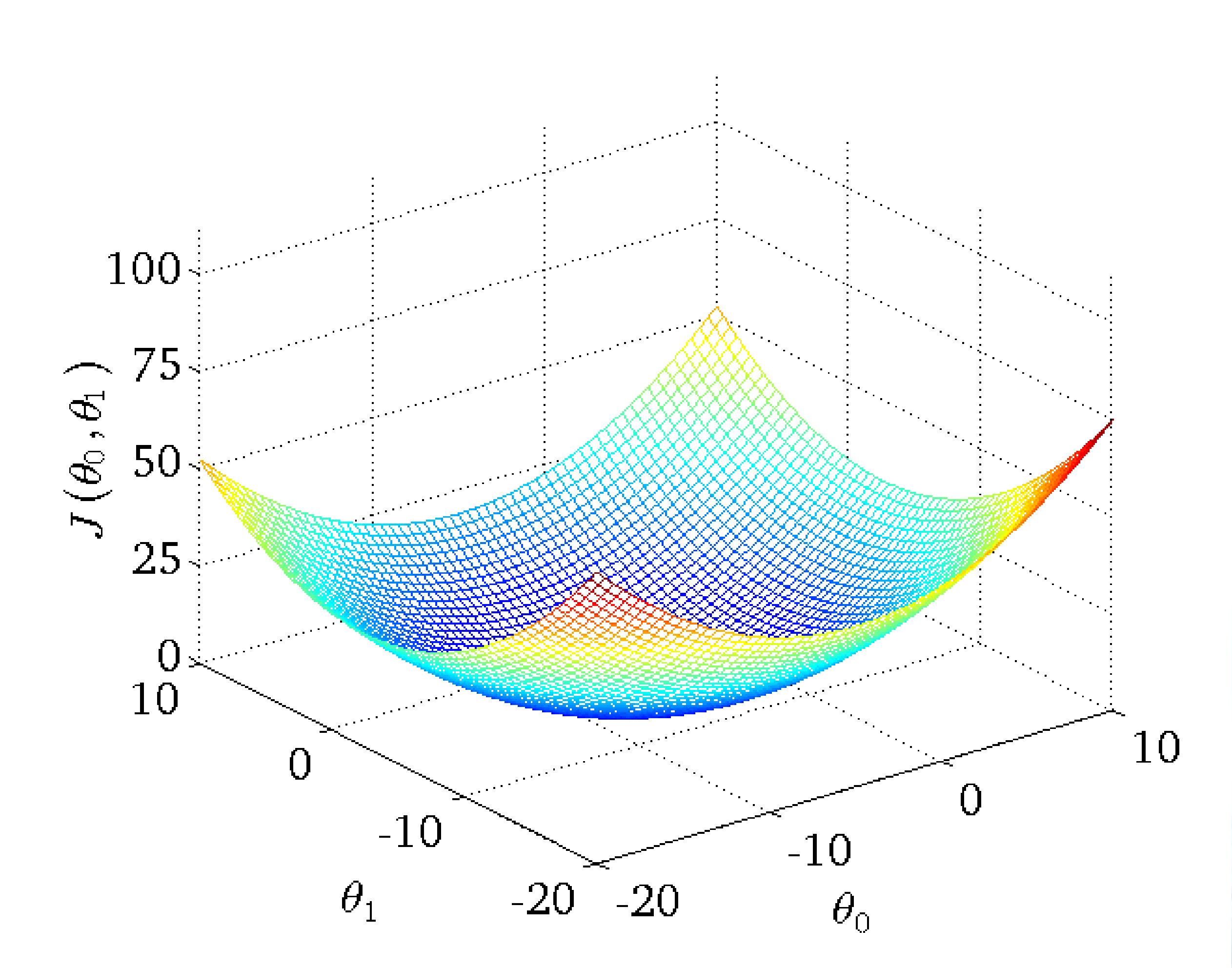


Gradient Descent



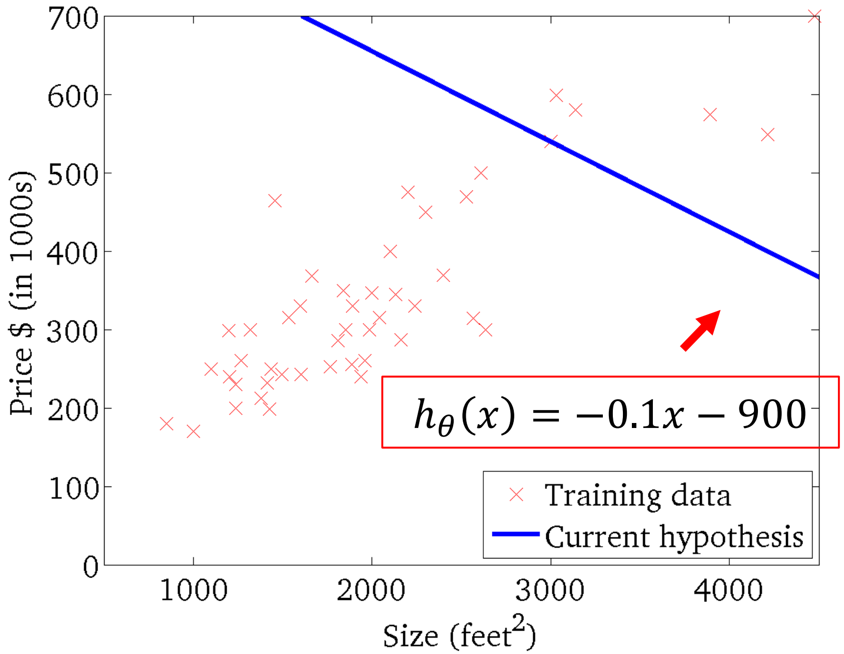
Gradient Descent for Linear Regression

- Convex function

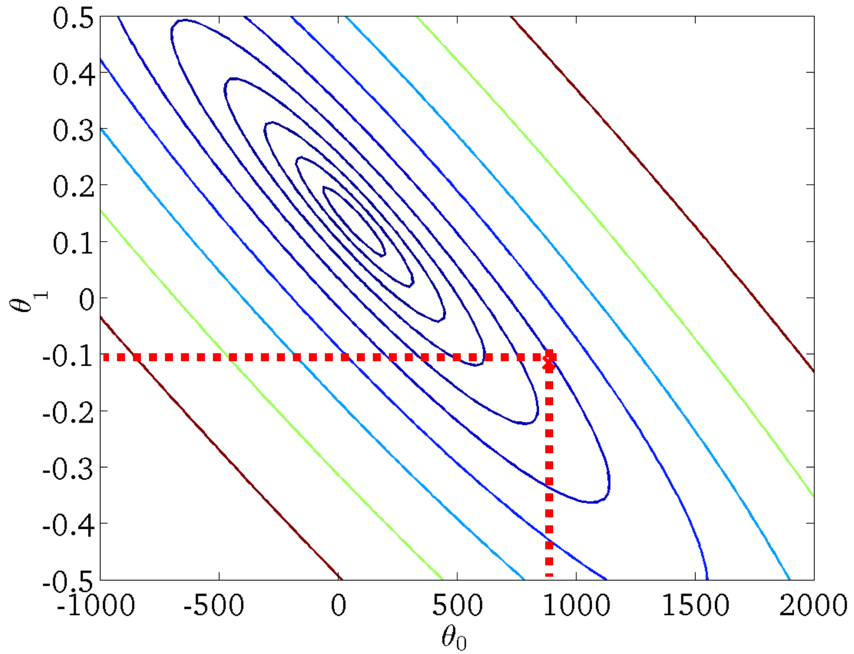


Gradient Descent for Linear Regression

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1

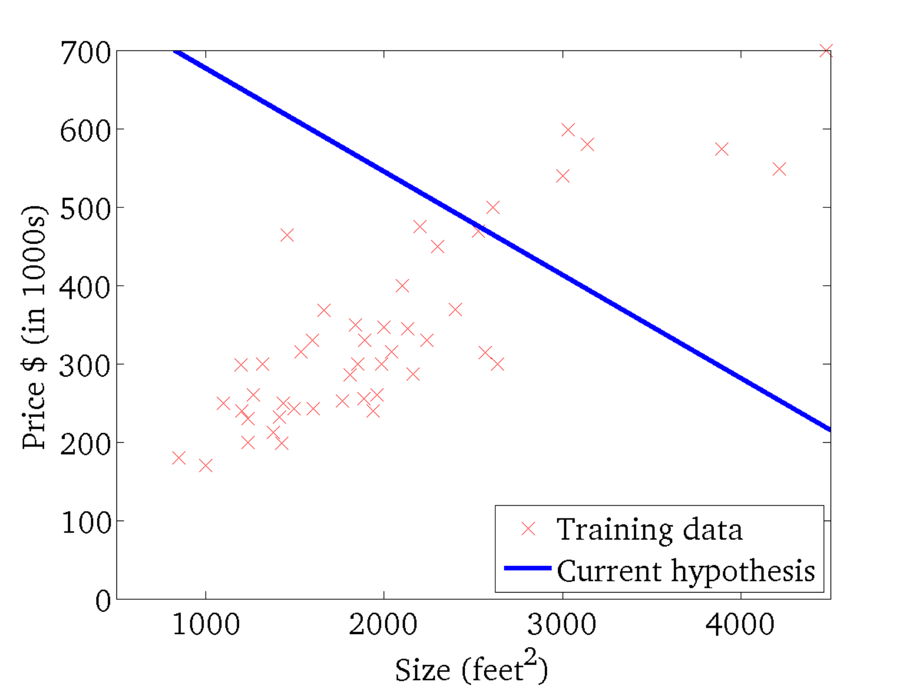


- $J(\theta_0, \theta_1)$
  - fct of the parameter $\theta_0,\ \theta_1$

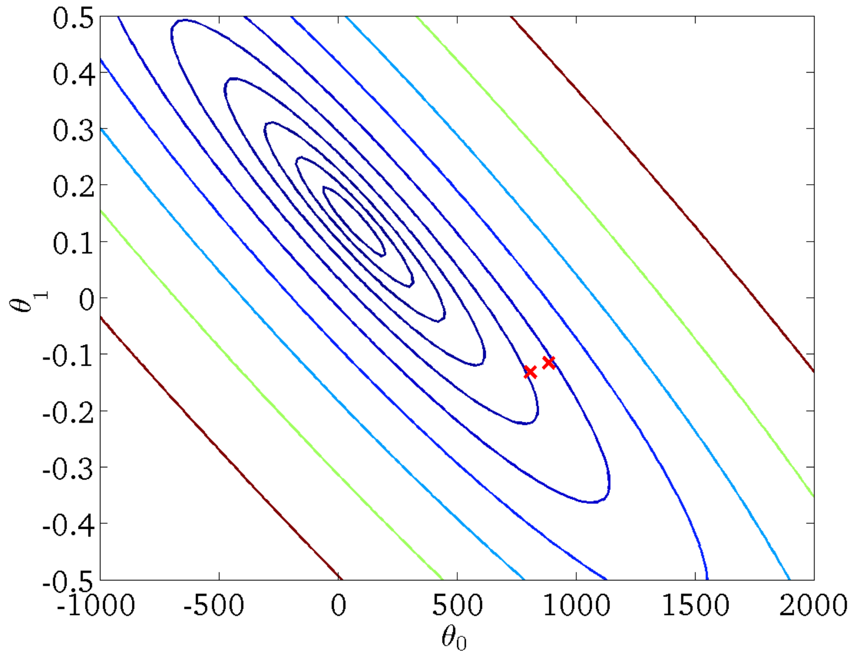


Gradient Descent for Linear Regression

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1

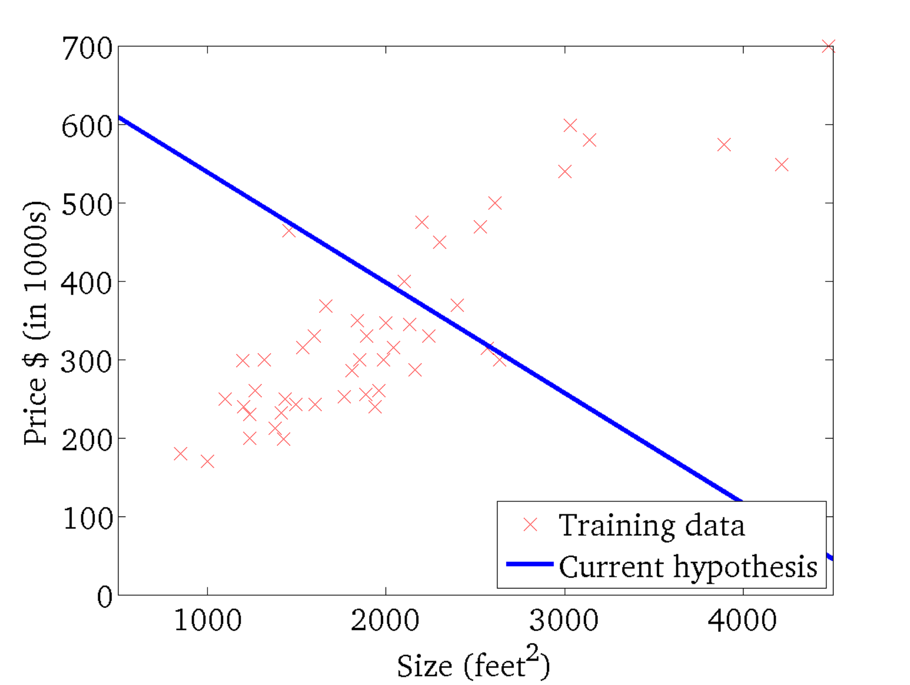


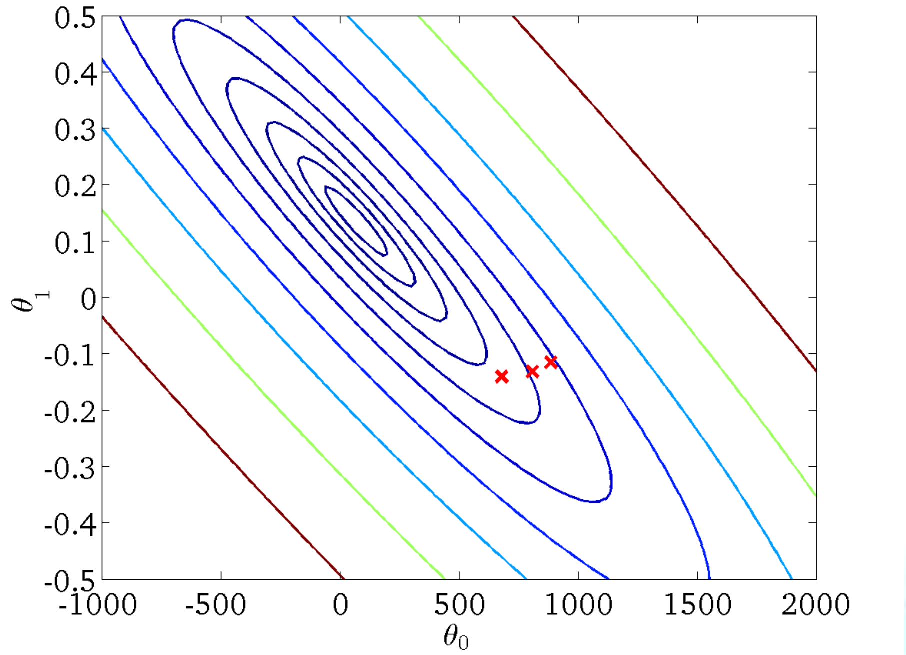
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1



Gradient Descent for Linear Regression

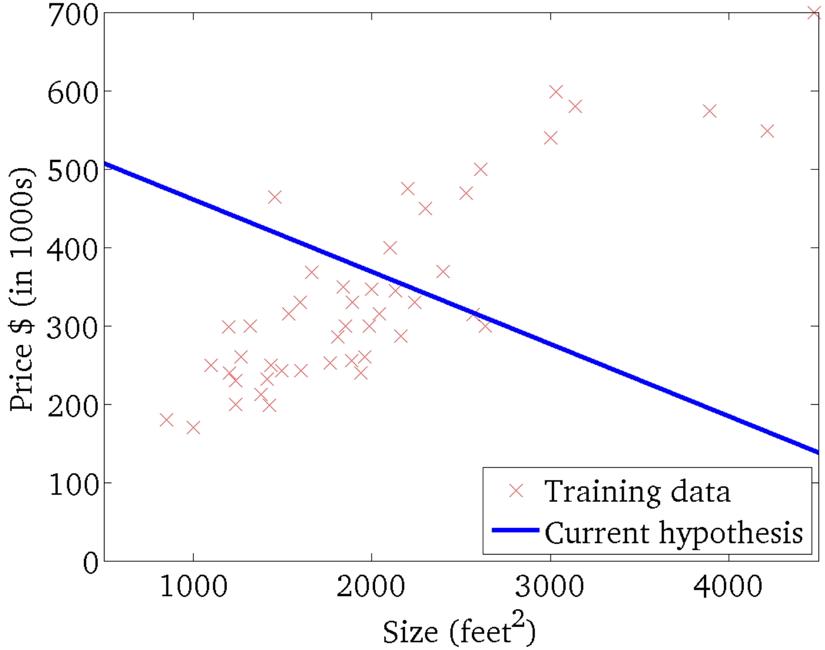
- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1
- $J(\theta_0, \theta_1)$
  - fct of the parameter $\theta_0,\ \theta_1$

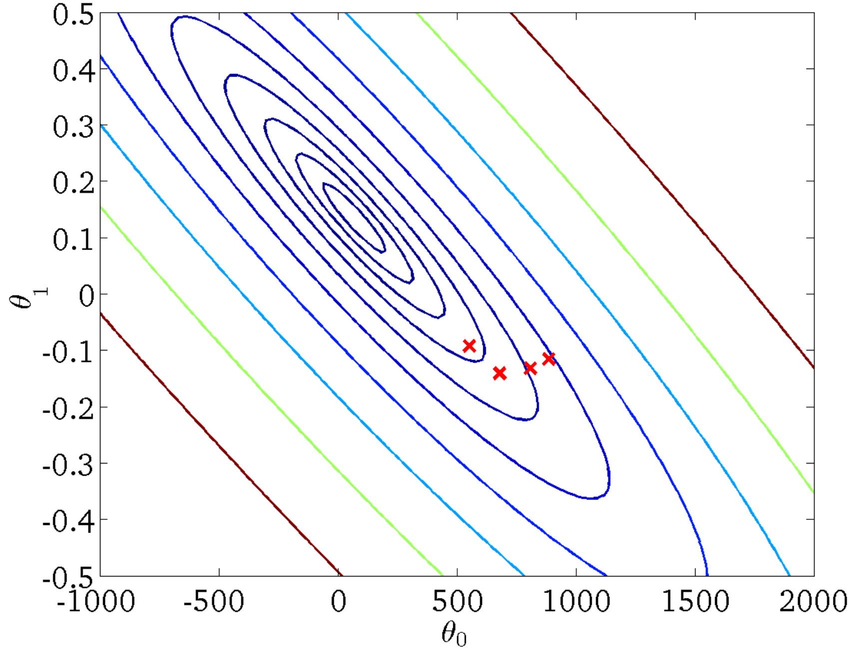




Gradient Descent for Linear Regression

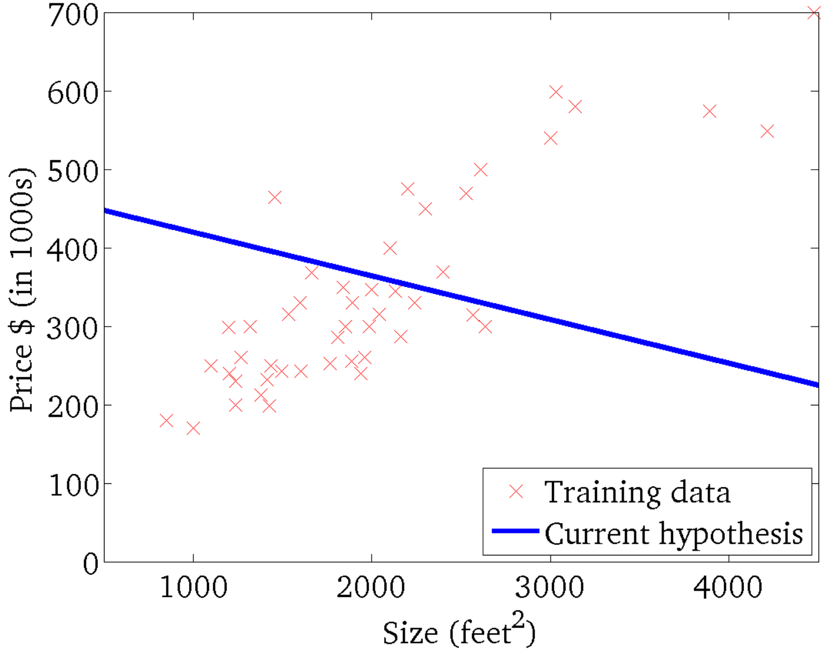
- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1

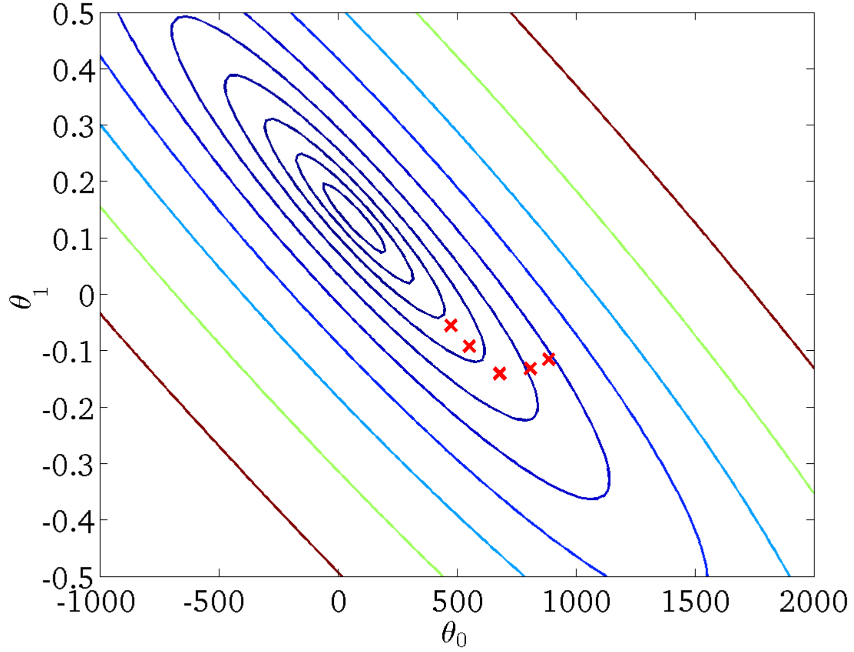




Gradient Descent for Linear Regression

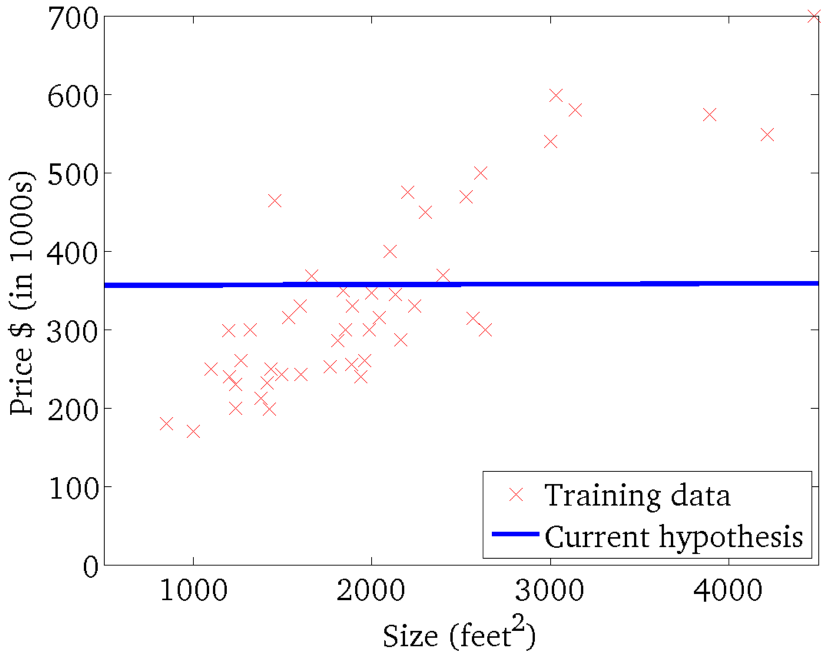
- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1

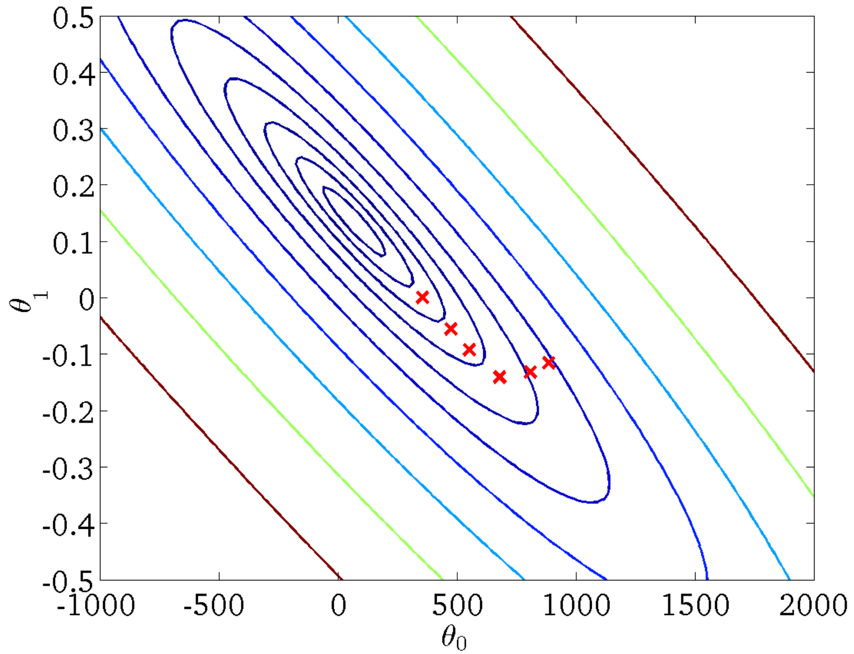




Gradient Descent for Linear Regression

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1

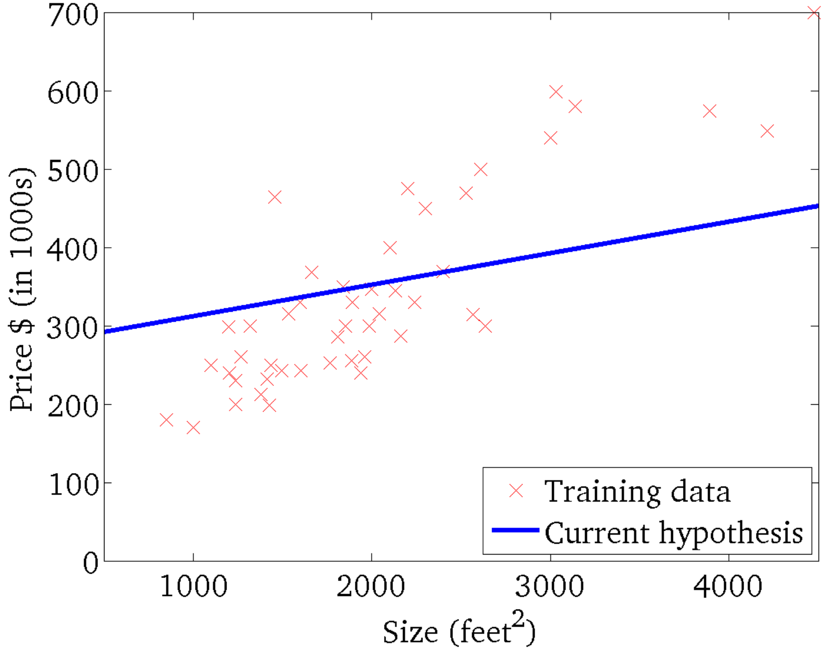


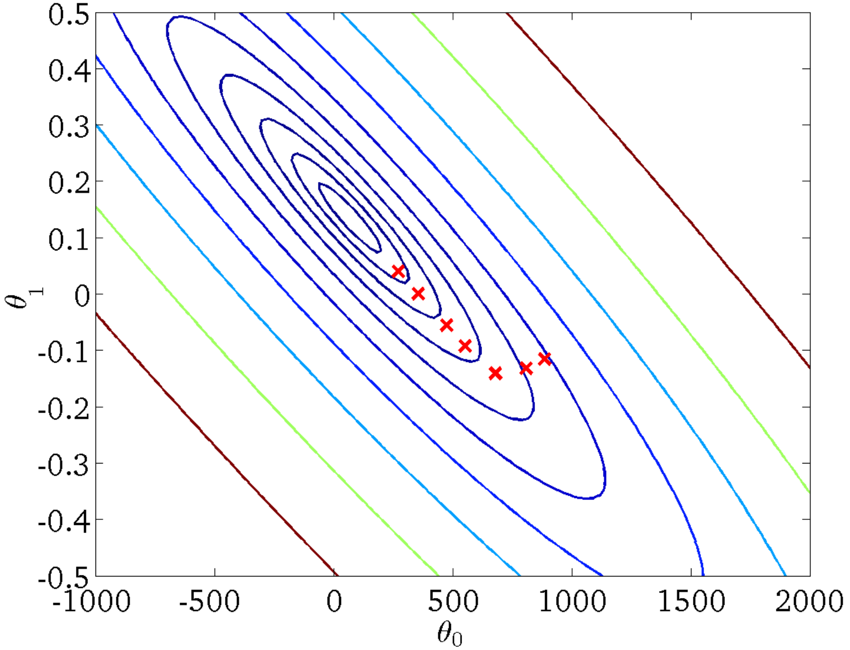


Gradient Descent for Linear Regression

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1

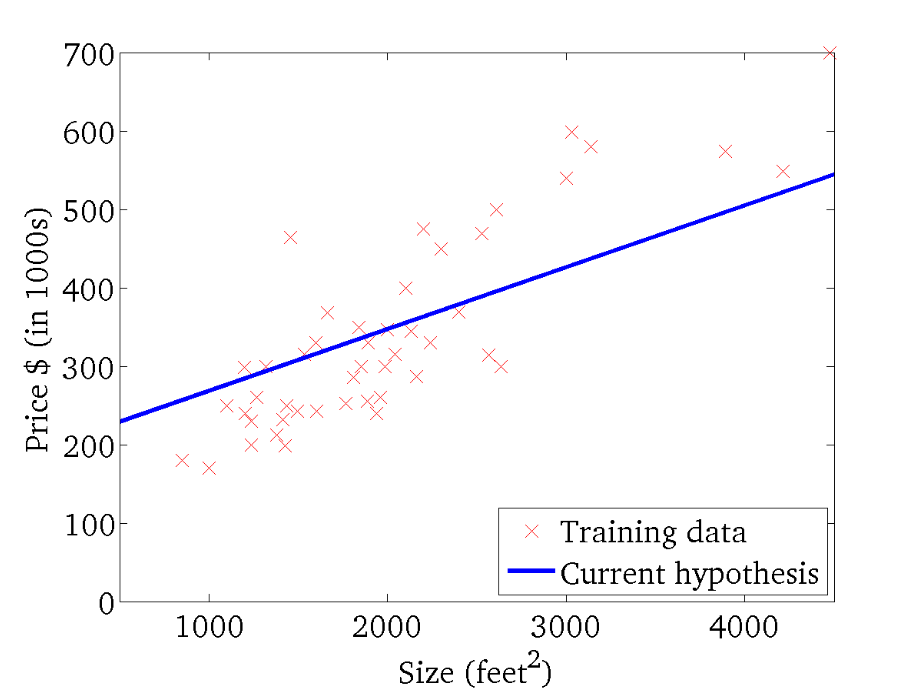
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1

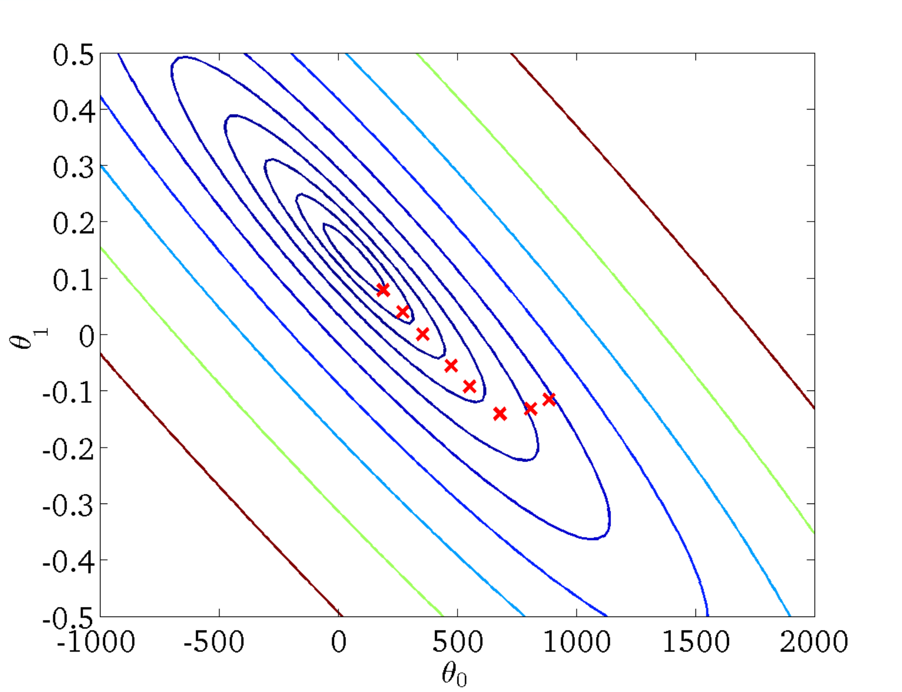




Gradient Descent for Linear Regression

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1
- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1

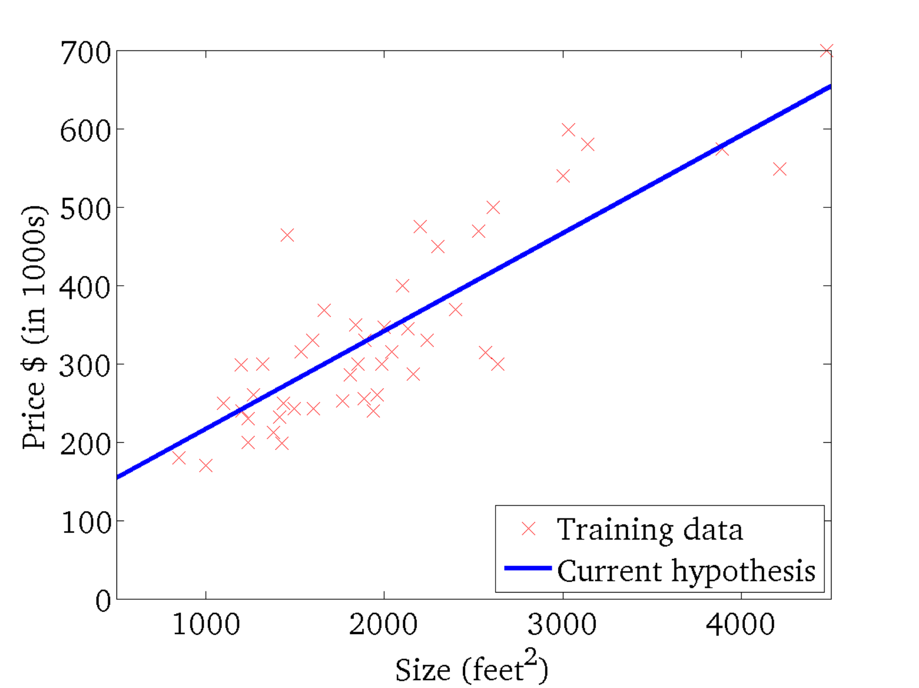


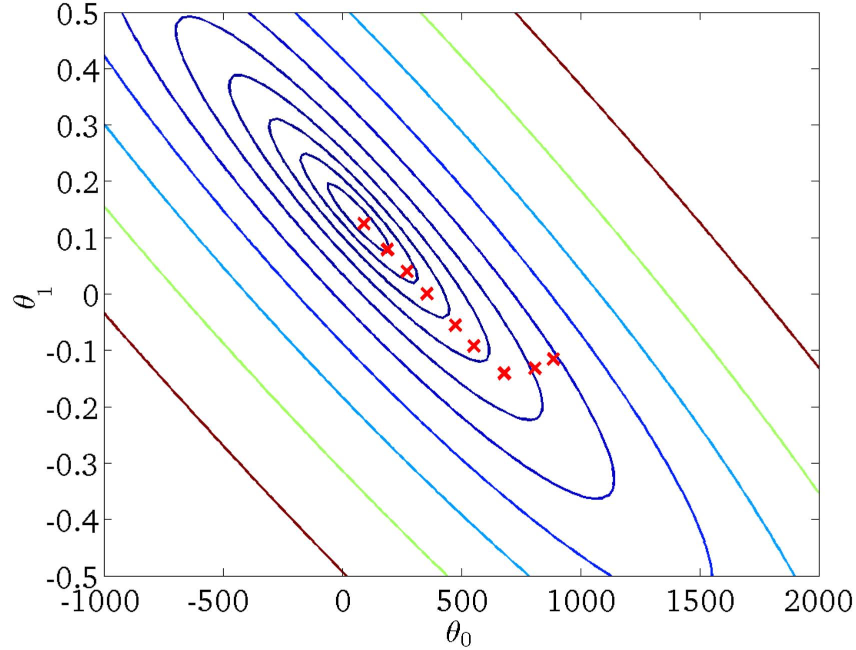


Gradient Descent for Linear Regression

- $h_\theta(x)$
  - a fct of x for fixed θ_0, θ_1

- $J(\theta_0, \theta_1)$
  - fct of the parameter θ_0, θ_1





"Batch" Gradient Descent

- "Batch"
  - Each step of gradient descent uses all the training examples.

- Repeat until convergence {

$$\theta_0 \leftarrow \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$\theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$$

}

(Update for θ_0 and θ_1 simultaneously)

References

- Andrew Ng, https://www.coursera.org/learn/machine-learning
- http://www.holehouse.org/mlclass/01_02_Introduction_regression_analysis_and_gr.html